ENHANCING EMPIRICAL ACCOUNTING MODELS WITH TEXTUAL INFORMATION

BY KHRYSTYNA BOCHKAY

A dissertation submitted to the

Graduate School—Newark

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Dr. Carolyn B. Levine

and approved by

Dr. Carolyn B. Levine

Dr. Alexander Kogan

Dr. Glenn Shafer

Dr. Miklos Vasarhelyi

Dr. Vladimir Vovk

Newark, New Jersey

May, 2014

ABSTRACT OF THE DISSERTATION

Enhancing Empirical Accounting Models with Textual Information

by Khrystyna Bochkay

Dissertation Director: Dr. Carolyn B. Levine

Rapid developments in information technologies and the increased availability of narrative disclosures in electronic form have provoked interest in textual analysis. In this dissertation, we survey research on textual analysis of mandatory and voluntary disclosures, describe methodologies for analyzing and incorporating text into quantitative models, and provide an analysis of MD&A text and earnings. Most empirical studies examine the association between text characteristics (e.g., tone and linguistic complexity) and future firm performance or market reactions. However, in-sample explanatory power is not equivalent to out-of-sample predictive power (Shmueli, 2010). We use regularized regression methods to examine whether textual disclosures in the Management Discussion and Analysis (MD&A) section of the 10-K report are helpful in predicting future earnings above and beyond traditional financial factors.

We develop techniques to combine textual information from the MD&A section of the annual report with financial variables and generate explicit firm-level forecasts of future earnings. We employ the "bag-of-words" (BOW) approach to represent MD&A

sections numerically and regularized regression methods to overcome problems of high-dimensionality and multicollinearity of data. We estimate and earnings forecasting models based solely on quantitative factors and compare them with models that include both quantitative information from financial statements and textual information from MD&A disclosures. We find that text-enhanced models are more accurate than models using quantitative financial variables alone. This supports the notion that the MD&A section has predictive value, one of the primary characteristics of relevance. Firms with larger changes in future performance, negative changes in future performance, higher accruals, greater market capitalization, and lower Z-scores have more informative MD&As, suggesting that MD&A content helps to reduce uncertainty. The MD&A is more informative in the period following recent regulatory reforms but less informative in the period covering the recent financial crisis, suggesting that managers may be unable to provide a reliable analysis of the business of the company in unstable economic periods. Finally, we show that financial analysts lose their forecasting superiority over text-enhanced statistical models for smaller firms and those with lower analyst following.

Acknowledgements

I would like to start this page with a quote of William Arthur Ward:
"*The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires.*"

This dissertation would not have been possible without the guidance of my exceptional committee members. You inspired me every day of my PhD life. You taught me what it takes to be an academic. You are and will be my role models.

Above all, I would like to thank my adviser Dr. Carolyn Levine for her continuous help, support, and patience. You were ALWAYS there and your optimism and enthusiasm I will never forget.

I am also grateful to Dr. Alexander Kogan for patiently answering all my questions and giving me good academic and life advice.

I would like to thank Dr. Glenn Shafer and Dr. Vladimir Vovk for believing in me and helping me to start my academic career.

I would like to thank Dr. Miklos Vasarhelyi for his continuous support and help with my research.

Special thanks to Dr. Michael Alles and Dr. Valentin Dimitrov. You taught me how to answer questions like "So, What?", "What is the theory behind all this?", etc. You encouraged me to work harder. I will never forget you support during the job search process.

Special thanks to Dr. Dan Palmon. Your jokes always made me smile :) You helped me to become a teacher.

Many thanks to all other faculty members and PhD students. Your advice and friendship

will be never forgotten.

Last, but by no means least, I would like to thank my loving husband Roman for his patience and support. You are my rock! My dearest parents, your Skype calls, love, and care motivated me every day.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

"One cannot guess how a word functions. One has to look at its use and learn from that."

– Ludwig Wittgenstein, philosopher

Business data have started being collected and accumulated at a dramatic pace over the past decade. With the availability of new technologies and low storage costs, businesses and regulators, among others, have started acquiring unbelievably vast amounts of digital information including emails, customer opinions and complaints, every-day transactions, various reports, news feeds, discussions, etc. For example, a retail giant Wal-Mart collects more than a million of customer transactions every hour; the Securities and Exchange Commission (SEC) collects all mandated reports from companies and makes them publicly available through an electronic data-gathering system EDGAR; Seeking Alpha, a free investor-oriented social media website, collects data from money managers, financial experts, market blogs, and investment newsletters and publishes more than two hundred market related articles daily. James Cortada who is the author of many books on the history of information in society states that "we are at a different period because of so much information" (see Economist (2010)). These examples emphasize that now we can do things that we previously could not: learn new business trends, strengthen decision-making, better exploit uncertainties, improve customer experience, share opinions, provide fresh insights on problems and so on.

The changes in the world of information have led to significant changes in both research and practice in accounting and finance. Now accounting and finance professions have started incorporating much more information while assisting investors, customers, auditors, creditors and other market participants. This information includes but is not limited to analysts' reports, conference calls with management, SEC mandated financial disclosures, financial news, social media and other linguistic descriptions of firms' current and future profit-generating activities. In this dissertation, we briefly introduce textual analysis techniques and tools currently used to conduct the linguistic analysis of narrative disclosures. Further, we discuss the trends in the accounting and finance disclosure literature and cover recent large sample studies that focus on the association of textual information with future performance and market reactions. Finally, we develop a methodology to combine textual information from narrative disclosures with quantitative information from financial statements to generate explicit firm-level forecasts of future earnings.

Over the history of accounting and finance research, academic studies have thoroughly examined the content of financial quantitative information (i.e., accounting numbers and market data), and how investors use that information to assess the health of companies (see for e.g., Ball and Brown (1968); Ou and Penman (1989); Fama and French (1993); Bushman and Smith (2001); Kothari (2001), among many others). Although all these studies have significantly enriched our general understanding of capital markets, they have also provoked some doubts about the underlying value and predictive ability of quantitative factors. Researchers have realized that incorporating information conveyed by quantitative factors alone may not be adequate to explain and/or predict the movement of stock prices, future firm performance, probability of default, cost of capital, etc. For instance, Shiller (1981), Roll (1988), Cutler et al. (1989), Lev and Thiagarajan (1993) show that the movement of stock prices cannot be explained by financial quantitative measures alone. Therefore, there is no reason to believe that market participants incorporate only quantitative information while making their decisions.

Most studies in the literature focus only on quantitative information for several reasons.

First, quantitative data can be easily accessed and downloaded from commercial databases (such as COMPUSTAT, CRSP, IBES, etc.). Second, quantitative data is more structured and objective. Financial statements, for example, are prepared in accordance with Generally Accepted Accounting Principles (GAAP) and we know how each item is measured and what it represents. In contrast, there is plenty of leeway with respect to qualitative information.[1] Even though some of the disclosures are mandated, managers still have a wide discretion about making their qualitative statements. For example, Management Discussion and Analysis (MD&A) is a required section of annual (10-K) and quarterly (10-Q) reports of all publicly traded companies. Managers use this section to provide the context for financial statements and allow investors to 'see the company through the eyes of management' (see SEC Release Nos. 33-8350; 34-48960; FR-72). MD&A disclosures are regulated by the SEC and there are numerous subjects that have to be addressed in the section. However, managers are not limited or guided by the SEC in expressing their own opinions, beliefs, and expectations, especially in forward-looking statements. Subjective nature of qualitative information and unavailability of such information in electronic form have imposed challenges on researchers in earlier years. Research designs were limited due to small sample sizes, hand-collected and hand-coded data. In contrast, today we have access to voluminous amounts of qualitative disclosures in electronic form and advanced technologies to use this type of data in our analyses.

Understanding the information value of qualitative disclosures is important for several reasons. First, companies regularly provide qualitative information to investors, creditors and others through SEC disclosures, explanatory statements, earnings calls, elaborations, financial media etc. An anecdotal evidence suggests that most of the information that flows out of corporations is qualitative in nature (Gangolly and Wu, 2000). Second, qualitative disclosures supply the context for reported financial numbers, i.e., they help investors

[1] We use the term *qualitative information* here to denote non-numerical disclosures, such as mandatory and voluntary linguistic disclosures. In contrast, *quantitative information* is numerical in nature (for example, earnings, market prices, analysts' forecasts, etc.)

understand the quality and variability of earnings and cash flows, future trends, and uncertainties. For example, managers may use qualitative disclosures to explain their choices regarding the accrual component of earnings or the value of intangible assets; this directly impacts the interpretation of financial statements. Third, managers' disclosure style may reveal some important management characteristics and help explain corporate decisions and strategies. A large body of literature in behavioral finance and economics suggests that people are prone to behavioral biases. Linguistic disclosures by managers can be used to estimate the impact of biases on professional management. Finally, if analysts and accounting variables are incomplete or biased measures of firms' fundamentals, qualitative disclosures may have incremental explanatory power for firms' future performance and returns (Tetlock et al., 2008). Therefore, analyzing the nature and impact of these qualitative disclosures can only improve our current understanding of the capital markets.

Obviously there are many challenges in studying the impact of qualitative disclosures. Qualitative disclosures do not have a clear structure, they are very subjective, and often ambiguous. With hundreds of words in English, there is an infinite number ways to communicate the same information (e.g., sales increased, sales improved, sales jumped, sales were higher, sales did not decrease, etc.). Although there are many disclosure standards and requirements, the structure of disclosures varies a lot across companies, industries, and markets. For instance, disclosures of more complex companies may be more involved, whereas they may be more straightforward for less complex companies.[2] Given the subjective and mostly unstructured nature of qualitative communications, it is difficult to find an objective methodology to extract useful knowledge from these outlets.[3]

Recent developments in text mining, statistics, machine learning, and computational

[2] In its MD&A requirements, the SEC states: "As the complexity of business structures and financial transactions increase, and as the activities undertaken by companies become more diverse, it is increasingly important for companies to focus their MD&A on material information" (see SEC Release Nos. 33-8350; 34-48960; FR-72). This suggests that more complex matters are more difficult to communicate.

[3] *Unstructured information* here refers to information that does not have explicit semantics ("structure") needed for computers to interpret the information (e.g., verbal document, email, speech, etc.). Unstructured information may be contrasted with the information stored in fielded form in databases or annotated (semantically tagged) in documents.

linguistics make it now possible to quantify the information conveyed in a large cross-section of unstructured verbal disclosures (see for e.g., Tibshirani (1996); Jurafsky and James (2000); Fan et al. (2006); Taboada et al. (2011), among others). It is now possible to develop computer programs that 'read' texts, 'hear' vocal communications, summarize documents, check redundancies, determine the tone or readability of texts, etc. These developments provide accounting and finance researches with useful tools to collect, process, and extract meaningful content from a growing body of corporate linguistic disclosures these days (Core, 2001).

Narrative disclosures are important because they help to bridge the gap between reported financial numbers and firms' economic environment. Commissioner Glassman in her speech on the quality of qualitative disclosures states that "simply complying with GAAP may leave gaps in disclosure and give investors an incomplete - or even misleading - picture of a company's operations."[4] Studies analyzing narrative disclosures focus only on the association between text characteristics (e.g., amount of disclosure, tone of disclosure, readability or complexity of disclosure) and future earnings or returns. These studies contribute to our general understanding of the information content and value relevance of qualitative disclosures. However, saying that certain variables are important in the explanatory setting is not the same as saying that those variables work well in the predictive setting. Two statistical constructs complement each other rather than substitute each other. Therefore, it is not clear whether we can build more accurate prediction models with the help of textual information.

Markets care about earnings. Earnings is arguably the most studied number from financial statements and forecasting earnings is an important task that has been the center of research for years now (see for e.g., Ball and Brown (1968), Fairfield et al. (1996), Gerakos and Gramacy (2013), among many others). For instance, earnings forecasts are used to value a firm, estimate implied costs of capital, determine marginal tax rates, and

[4]See *www.sec.gov/news/speech/spch041003cag.htm*.

explain market anomalies. Papers looking at earnings forecasting use only quantitative information, ignoring the fact that most of the disclosed information is qualitative (textual) in nature. Alternatively, researchers use analysts' consensus forecasts as a proxy for earnings expectations, relying on the belief that the combination of analysts' expertise and access to multiple sources of information results in superior forecasts of future earnings. However, given that analysts are not paid by investors, their earnings forecasts and investment recommendations may often be biased. Therefore, it is important to develop independent (bias-free) models for earnings prediction. Moreover, independently constructed time series models allow researchers to generate earnings forecasts for firms that do not have analyst following or lack long time series of earnings realizations.

This dissertation sets out to predict earnings using both quantitative information from financial statements and qualitative information contained in Management Discussion and Analysis (MD&A) section of annual (10-K) report. We develop techniques to represent textual information numerically and combine it with accounting-based numerical variables. We introduce two ways of incorporating text into a forecasting model: one based on text categories and one based on a more detailed word-by-word analysis. The advantage of the former approach is simplicity and low-dimensionality: frequencies of words are aggregated into one predefined category. However, its major disadvantage is the loss of information through aggregation. The latter approach effectively addresses the problem of extreme aggregation by allowing a statistical model to determine the relevance of individual words. The detailed word-by-word analysis captures the predictive content of textual information better at cost of increasing the number of predictor variables. We introduce several methods to deal with high-dimensionality and multicollinearity of textual data, including feature selection methods and regularized regression methods such as Ridge Regression, Kernel Ridge Regression, and Lasso Regression. We find that the detailed textual analysis method is superior to simple text aggregation. Using this method, we also find that the textual content of the MD&A section is helpful in predicting future earnings, above and

beyond traditional financial factors. Forecasting models based on quantitative and qualitative information are both statistically and economically superior to those that use only quantitative information. We also find that the predictive value of MD&A texts varies with firm characteristics. MD&A section is more informative for firms with larger changes in future performance, negative changes in future performance, higher accruals, greater market capitalization, and lower Z-scores. These findings suggests that MD&A information provides content for financial statements and helps to reduce uncertainty. In addition, we find that MD&A is more informative in the period following recent regulatory reforms but less informative in the period covering the recent financial crisis. Finally, we compare the accuracy of text-enhanced time series forecasts to the accuracy of analysts' consensus forecasts and find that financial analysts lose their forecasting superiority over text-enhanced statistical models for smaller firms and those with lower analyst following.

The rest of the dissertation is organized as follows. Chapter 2 provides an extensive survey of the literature that uses textual analysis to examine the information content of narrative disclosures. Chapter 3 introduces research questions and hypotheses. Chapter 4 develops techniques to combine textual information with quantitative information for earnings prediction. Chapter 5 reports the results. Finally, Chapter 6 concludes the dissertation.

Chapter 2

Literature Review

2.1 Textual Analysis: Research Methodologies

Textual analysis is an empirically grounded method used to process, extract, and interpret the characteristics of natural language texts. The fundamental problem of textual analysis is to represent texts numerically, so they can be used as inputs in statistical models. Text mining techniques are all based on counting words, phrases, or other verbal elements embedded in texts. Conceptually, there are at least three text-based disclosure measures interesting to researchers: the amount, the tone (positive or negative), and the transparency (or readability) of disclosure (Li, 2010). In this chapter, we briefly overview the methodologies of automated textual analysis. Manning and Schütze (1999), Berry and Castellanos (2004) and Liu (2012) provide a more detailed summary of textual analysis methods and techniques.

Content Extraction

There are two main approaches to the problem of automatically extracting the content from a large number of qualitative disclosures: dictionary-based and text-classification. The dictionary-based approach involves counting words in a document and classifying them into categories based on a predetermined dictionary. For instance, the tone of a document (positive or negative) can be determined by counting positive and negative words

from an existing dictionary that occur in the document. The text-classification (or statistical approach) involves building text classifiers from labeled elements in documents (Pang et al., 2002). The text-classification approach is essentially a supervised machine learning method, i.e., an algorithm is first trained on a subset of documents that have labels and then used to classify unlabeled documents.

There are advantages and disadvantages of both text analysis methods and there is no uniform answer to what method works best with corporate disclosures. One of the disadvantages of the dictionary-based approach is that the number of dictionaries is relatively small. There are two dictionaries used in most accounting and finance studies to determine the tone of disclosures: Harvard's General Inquirer dictionary and Loughran and McDonald (2011)'s financial sentiment dictionary. Results of textual analysis using the dictionary-based approach depend significantly on the dictionary domain specificity. General Inquirer, developed for psychology and sociology, is not a business-oriented dictionary and some of its words do not translate well into the realm of corporate disclosures[1]. In contrast, Loughran and McDonald (2011)'s dictionary is developed using a large number of SEC annual reports over the period 1994-2008. Many studies have recognized the benefits of using domain-specific dictionaries to determine the tone of disclosures (see for e.g, Feldman et al. (2009); Henry and Leone (2010); Loughran and McDonald (2011)). Positive features of the dictionary-based textual analysis include lower computational costs (i.e., fixed number of words to analyze) and more objective word analysis (i.e., the composition of the dictionary is beyond the control of the researcher).

The statistical text classification starts with creating labeled training data (usually with the help of human raters). To determined the tone of each text element (word, phrase, sentence, paragraph, or the whole document) in the training set, human raters read the text element and decide whether it is positive, negative, or neutral. Then, a computer algorithm

[1]Loughran and McDonald (2011) provide an excellent discussion about domain specificity in the dictionary-based text analysis.

is trained using the labeled text and tested using the new previously unseen data. Statistical classification methods reach quite high accuracy in detecting the tone of a document (Chaovalit and Zhou, 2005; Boiy et al., 2007). However, although such methods perform very well on the documents that they are trained on, their performance drops dramatically (to almost to chance) when the same algorithm is used on different documents (Aue and Gamon, 2005). This makes the statistical text classification methodology more costly and less applicable to different disclosure types. One of the benefits of this approach is that an algorithm incorporates the content of a sentence better. For example, if a sentence is about expenses, then the word "increase" would be coded as negative even though it is a positive word by itself (see Li (2010)). In Section 2.2, we discuss accounting and finance research studies that use dictionary-based and statistical text classification approaches for content analysis. In Section 4.1.2, we explain the numerical representation of texts using word counts in greater detail.

Text-based Measures

While working with qualitative disclosures, researchers are generally interested in developing one measure for the text in interest. There are at least four text-based measures that can be calculated using the disclosure content. These measures are:

(a) *The amount of disclosure*. This measure is typically based on the total number of words (a proxy for disclosure length or size) used to describe a certain topic. For example, to estimate the extent to which managers focus on risk in their annual reports, we can simply count the number of words in all sentences that mention words "risk, risks, risky, etc." Alternatively, we can count the number of sentences, assuming that the length of the sentence does not vary across the document.

(b) *The tone of disclosure*. To calculate the tone of a document, we can count the number of occurrences of positive and negative words in the document and then aggregate these counts into a single score for the document. Under this approach, we need to have a dictionary of positive and negative words. Alternatively, we can build an

algorithm to classify each document as positive or negative (see our discussion of the content extraction methods above).

(c) *The readability of disclosure*. The readability of disclosure or disclosure transparency is the measure of how easy it is to read a document. Many papers in the literature use the Fog Index to estimate the readability of texts. The Fog index is based on the numbers of words per sentence and the proportion of complex words in the document. The idea is that reading more complex words and longer sentences requires more effort.

(c) *The similarity of disclosure*. Document similarity measures can be used to determine how different two documents are. These measures can be used to compare documents of the same company over time or to compare documents of different companies at the same time. In text mining, one of the most popular document similarity measures is the cosine similarity that measures the difference between two vectors of word frequencies (Brown and Tucker, 2011).

2.2 Textual Analysis of SEC filings

In this section, we discuss accounting and finance studies on textual analysis of the SEC disclosures. Our particular interest is on the Management Discussion and Analysis Section of annual (10-K) and quarterly (10-Q) reports.

Although there are many difficulties in studying the impact of qualitative factors on financial markets, current accounting and finance researchers seem to be very interested in this topic. This interest is caused by availability of qualitative disclosures in electronic form and development of different computerized methods for textual analysis. The SEC's filing system EDGAR, websites and databases with transcripts of earnings conference calls, analysts' reports, financial news, among other sources, provide researchers with rich resources to test different economic theories. In addition, significant developments in the fields of

computational linguistics and machine learning make the analysis of large textual datasets possible.

Qualitative disclosures continue to be an integral part of managers' communication with investors. Since the 1980s, the Securities and Exchange Commission (SEC) has required that companies include qualitative disclosures in annual (10-K) and quarterly (10-Q) financial reports in a separate section called Management Discussion and Analysis (MD&A). According to the SEC, the principal objectives of the MD&A section are: "(1) to provide a narrative explanation of a company's financial statements that enables investors to see the company through the eyes of management; (2) to enhance the overall financial disclosure and provide the context within which financial information should be analyzed; and (3) to provide information about the quality of, and potential variability of, a company's earnings and cash flow, so that investors can ascertain the likelihood that past performance is indicative of future performance" (see SEC Release No. 33-8350).

Whether SEC qualitative disclosures contain incremental information is an empirical question. Given so many regulatory reforms and requirements, one may argue that it obvious that MD&A section (or the whole 10K(Q) report) is informative. However, the SEC often raises concerns regarding boilerplate redundant disclosures and cautions firms against misleading the reader. Managers may be reluctant to provide information in the SEC filings because of concerns over proprietary costs. In addition, the MD&A section is not audited and its forward-looking statements (i.e., statements that are arguably the most informative) are protected by safe harbor provisions. Therefore, it is not so obvious that SEC disclosures are informative.

To improve our understanding of information contained in qualitative disclosures, many accounting and finance studies analyze different characteristics of mandated disclosures. While early research studies on MD&A are mostly descriptive in nature, more recent studies focus on the content, characteristics, and information value of MD&A. For example, Bagby et al. (1988) give a wide overview of SEC and other qualitative disclosures. Using a critical examination of legal cases related to mandated disclosures, the study analyzes

the usefulness of SEC disclosures within a wider class of federal disclosures. Dieter and Sandefur (1989) focus on MD&A mandated requirements and provide guidelines on drafting an MD&A that would satisfy SEC regulations.

In 1989, the SEC issued a number of disclosure requirements that were expected to make the MD&A section more informative. Hooks and Moon (1993) develop a classification scheme for measuring the difference between expected and actual MD&A disclosure frequencies. The test runs support that the developed classification scheme is effective for analyzing disclosures and permits comparisons regarding the consistency of disclosure frequency with expected frequencies of underlying events. In addition, the study assesses corporate managements' response to the issuance of the SEC MD&A guidelines. The findings indicate that the companies appear to respond to the release of the SEC 1989 guidelines by increasing the level of their disclosures. The disclosure response is even stronger one year after the 1989 SEC reform.

One of the first papers that develops a methodology to statistically evaluate the content of narrative disclosures is Frazier et al. (1984). To evaluate textual information in annual reports, the authors develop a methodology which consists of a content analysis system called WORDS. WORDS helps to identify the most important words/factors that indicate positive or negative narrative themes for a sample of 74 annual corporate reports in 1978. The study finds that the positive and negative factors (and the associated themes) can predict the cumulative abnormal annual returns for the next year (1979 in their sample). In addition, the results indicate that are no significant differences in narrative disclosures across the ownership structure of the companies examined.

Pava and Epstein (1993) test whether the qualitative content in MD&A provide useful information about a company's future performance using reports and financial data of 25 randomly selected companies in 1989. The study finds that while most companies provide good explanations of historical events, very few provide useful forecasts for the future. Companies seem to have a strong bias in favor of correctly identifying positive news, whereas negative news are either omitted or unclearly reported. These results suggest that

either managers tend to be much better in forecasting favorable news rather than unfavorable or they tend to withhold bad news. In addition, the study finds that managers predict company-specific events better than either industry- or economy-specific events.

One of the SEC concerns regarding the mandated reports is the clarity of disclosed information. The SEC frequently emphasizes the importance of understandable to investors reports. Schroeder and Gibson (1990) is one of the earliest papers that studies the readability of the MD&A section and the president's letter. Readability is defined as "the quality in writing which results in quick and easy communication. Readable writing communicates precisely - and with a single reading" (Lesikar and Lyons, 1986). Using techniques from the psychology literature, the study constructs a measure of the reading ease which is based on the word and sentence length and the use of passive voice in sentences. Contrary to the SEC expectations, the results show that MD&A texts are, in general, difficult to read and comprehend.

There are several earlier studies in the literature that analyze the association between MD&A disclosures and future performance, stock returns, or analyst forecasts. For example, Bryan (1997) examines the association between MD&A disclosures and future financial variables such as the directions of changes in future sales, future earnings per share, future cash flows from operations, and future capital expenditures. Using a sample of 250 MD&A sections in 1990 (a year after SEC issued clearer guidelines), the study demonstrates that MD&A disclosures are significantly associated with one-year-ahead changes in sales, earnings per share, and capital expenditures, but not operating cash flows. In addition, Bryan (1997) shows that MD&A disclosures are positively and significantly associated with financial analyst forecasts around the release date of MD&As. Controlling for financial statement information, the study also finds that the capital expenditure disclosure is significantly associated with current and future stock returns. Likewise, Callahan and Smith (2004) find that their disclosure index based on comprehensive content analysis of MD&A provides incremental explanatory power for future firm performance and market valuation.

Cole and Jones (2004) focus on MD&A disclosures of retail companies to show that certain types of disclosures (namely, disclosures related to sales growth, store openings and closings, and capital expenditures), can predict or explain future profitability and stock returns. Using data from 150 companies for the period 1996-1999, they find that disclosure-based variables can predict future revenues, earnings, and are associated with contemporaneous stock returns. These results suggest that MD&A disclosures are useful to market participants and should be encouraged.

Whether MD&A section is a part of a firm's overall disclosure package is the focus of Barron et al. (1999) and Clarkson et al. (1999). Barron et al. (1999) test the association between properties of analysts' earnings forecasts and MD&A quality, where MD&A quality is measured by compliance ratings assigned by the SEC personnel. Using a large sample of MD&A disclosures in 26 different industries and controlling for quantitative financial factors, the authors find that there is significant association between MD&A ratings and analysts' earnings forecasts. The results are driven by historical and forward-looking disclosures related to capital expenditures and operations. Relatedly, Clarkson et al. (1999) investigate the role of MD&A in a firm's disclosure package with questionnaires. They find that MD&A disclosures are useful to sell-side analysts who are members of the Toronto Society of Financial Analysts (TSFA). Additionally, based on a sample of 55 companies on the Toronto Stock Exchange in 1991 and 1992, they show that the quality of various subsections of the MD&A disclosures is generally influenced by expected firm performance, financing activities, firm size, independent press reports, and major firm related events.

The rapid development of information technologies in recent years has resulted in proliferation of papers that analyze high volumes of quantitative and qualitative data. Li (2008) is one of the recent papers on linguistic analysis of the SEC disclosures which analyzes the relationship between the readability of annual reports and earnings quality. The study is motivated by the SEC's plain English disclosure regulations that were released to make

disclosures more readable and understandable.[2] Li (2008) builds his analysis on the "management obfuscation hypothesis" which states that if markets underreact to information that is more convoluted then managers may have greater incentives to obfuscate information when firm performance is poor (Bloomfield, 2002). The "management obfuscation hypothesis" suggests that the earnings quality is lower when managers provide more convoluted (i.e., less readable) disclosures. Li (2008) hypothesizes that the positive earnings of firms with less readable annual reports are less persistent, while the negative earnings of such firms are more persistent. The readability of annual reports is measured by the Fog Index and the length of a document. The Fog Index, originated in computational linguistics, is constructed on the basis of syntactical textual features including words per sentence and syllables per word. The length of of document is measured by total words. Consistent with the hypothesis, Li (2008) finds that companies with low earnings tend to disclose information in a more "difficult-to-read" manner, whereas companies with high earnings prepare more readable reports.

Several studies try to analyze the effects of reporting complexity on market reactions. For instance, Lee (2012) tests whether less readable and longer quarterly (10-Q) reports affect market reaction to earnings. The study finds that stock prices of firms with longer or less readable 10-Qs react less strongly to the earnings-related information during the 3-day window following the 10-Q release. In addition, there is greater information asymmetry during the 10-Q release window for firms with less readable quarterly reports. These results suggest that the transparency of qualitative disclosures influences how the stock market processes earnings-related information. Likewise, You and Zhang (2009) examine investors' responses to information contained in annual (10-K) reports. Using a sample of 10-K reports filed during 1995-2005, the study documents unusual stock price and trading volume movements around the 10-K filing dates. There is a positive association between abnormal stock price movements and future accounting performance, suggesting that 10-K

[2]New SEC regulations require issuers to use the following standards while preparing their annual and quarterly reports: short sentences; definite, concrete everyday language; active voice; tabular presentation of complex information; no legal jargon; and no multiple negatives (see SEC (1998)).

report contains some useful information. In addition, the study finds that the information complexity of 10-K filings (as measured by word counts) causes stronger under-reaction of market participants.

Miller (2010) investigates the effects of reporting length and readability on small and large investors' trading behavior (volume and consensus) around the 10-K filing dates. Using a large number of 10-K reports issued between 1995 and 2006, the study finds that more complex reports are significantly associated with lower levels of aggregate trading volume. This relationship appears to be driven by a reduction in small investor trading volume. The results are consistent with the information processing cost hypothesis: more complex disclosures are too costly for small investors to process in the short window surrounding the filing date.

Whether sell-side financial analysts are influenced by the complexity of annual reports is the focus Lehavy et al. (2011). More specifically, Lehavy et al. (2011) examine how 10-K readability is associated with sell-side financial analyst following and the properties of their earnings forecasts. Using the Fog Index as the readability measure, the authors find that analyst following, the amount of effort incurred to generate their reports, and the informativeness of their reports are greater for firms with less readable 10-Ks. In addition, less readable 10-Ks are associated with greater forecast dispersion, lower accuracy, and higher overall uncertainty in analyst earnings forecasts. The results in the paper suggest that management communications lead to an increasing demand for analyst services and a greater collective effort by analysts.

Different proxies of document complexity and readability are also considered in the literature. For instance, Loughran and McDonald (2014) proxy for document complexity/readability with document file size. The proposed measure outperforms the widely-used Fog Index, it does not require document parsing and is easy to replicate. The measure is correlated with alternative readability measures. Controlling for financial variables, the study finds that larger 10-K file sizes have significantly higher post-filing date abnormal

return volatility, higher absolute standardized unexpected earnings (SUE), and higher analyst dispersion. Loughran and McDonald (2014) suggest that the SEC should focus less on style and instead encourage managers to write more concisely.

Several papers examine capital market anomalies, including the accrual anomaly and the post-earnings announcement drift, using the tone or sentiment (positive or negative, risky or risk-free) of qualitative disclosures. Li (2006) tests market efficiency with respect to the risk sentiment in texts of 10-K fillings. The study examines whether the risk sentiment and changes in the risk sentiment expressed in the annual reports are associated with future firm performance and stock returns. Using a large sample of annual reports, Li (2006) measures the risky tone of annual reports by counting the frequency of words related to risk or uncertainty in the whole 10-K document. The study finds that an increase in the number of risky words in annual reports is strongly associated with lower future earnings and stock returns. A hedge portfolio that is based on long positions in firms with small increase in risk sentiment and short positions in firms with a large increase in risk sentiment generates significant abnormal returns (more than 10%, excluding transaction costs). These results suggests that the stock market fails to fully incorporate the textual information in annual reports about future performance and stock returns.

It is always a concern that the dictionaries built by researchers from other fields (e.g., psychology) may not be appropriate for doing content analysis of financial texts (see for e.g., Henry and Leone (2010)). Loughran and McDonald (2011) effectively address this issue by developing a sentiment dictionary that is suitable for business language. Using a large sample of 10-K reports during 1994-2008, the authors create negative, positive, uncertainty, litigious, modal strong, and modal weak word lists and show that these word lists better reflect the tone in financial texts. Some of the word lists are related to market reactions around the 10-K filing date, trading volume, unexpected earnings, subsequent stock return volatility, and events such as accounting fraud or reported material weaknesses in accounting controls.

Feldman et al. (2009) examine the tone of the MD&A section and both short-term

and long-term stock price reactions. To measure the tone of management discussions, the authors classify words into positive and negative categories using financial sentiment dictionaries developed by Loughran and McDonald (2011). Feldman et al. (2009) find that the tone change is significantly associated with both short-window return around the filing date and the drift returns in the post-filing period, even after controlling for accruals and earnings surprises. The incremental value of management's tone change depends on the strength of the firm's information environment. These results suggests that management discussions are informative and the market under-reacts to this information.

The most recent study that refines the measurement of document tone is Jegadeesh and Wu (2013). Jegadeesh and Wu (2013) develop a new tone measurement methodology using market reactions to words in Loughran and McDonald (2011)'s financial sentiment dictionary. The proposed measure of document tone based on market weighting for 10-Ks is significantly associated with market returns of filing firms around the 10-K filing dates. The tone measure is strongly related to filing date returns for both positive and negative word lists, unlike findings in prior studies that only negative words matter (see for e.g., Loughran and McDonald (2011)). The measure of tone is significantly related to filing date returns even after controlling for financial factors such as earnings announcement date returns, accruals and volatility. Jegadeesh and Wu (2013)'s methodology to measure the tone of disclosures minimizes the level of subjectivity required for content analysis as weights of words and their tone are determined by the market, not by a small group of researchers.

There are several papers that analyze the association between the tone of SEC disclosures and past and future firm performance and market reactions. For example, Abrahamson and Amir (1996) use a computerized content analysis tools to measure the information content of 1,300 president's letters to shareholders written between 1986 and 1988. The results indicate that the information contained in the president's letters is significantly associated with financial performance measures (such as percentage change in sales, earnings levels divided by stock price, book rate of return, and dividend changes divided by stock

price). More specifically, the relative negative content of a letter is strongly negatively associated with past and future performance, strongly negatively associated with past and current annual returns, and weakly negatively associated with future returns. Overall, all the findings indicate that the information found in the president's letter is consistent with reported financial information.

More recently, Li (2010) examines the tone of the forward-looking statements in the MD&A section of 10-K and 10-Q filings. To calculate the tone of forward-looking texts, the study uses a Naive Bayesian machine learning algorithm on a large sample of 10-Ks and 10-Qs filed between 1994 and 2007. The results indicate that when managers are more optimistic in their forward-looking statements, future performance is better, suggesting that forward-looking statements are informative. In addition, Li (2010) shows that companies with better current performance, lower accruals, smaller size, lower market-to-book ratio, and less return volatility tend to have more positive forward-looking statements in MD&As. The tone of the MD&A section has not changed after recent regulatory changes by the SEC and the passage of the Sarbanes-Oxley Act.

Merkley (2013) examines whether earnings performance is related to firms' narrative R&D disclosure decisions. Using a large sample of R&D disclosures in annual reports, the study finds that current earnings performance is negatively associated with the quantity of narrative R&D disclosures. Analyzing the detail, tone, and readability of narrative R&D disclosures, Merkley (2013) also finds that managers use R&D disclosures to provide relevant information rather than obfuscate.

Whether MD&A disclosures provide a narrative explanation of a company's financial statements is the focus of Sun (2010). Sun (2010) investigates whether MD&A disclosures help users of financial information to interpret disproportionate inventory increases. The study identifies 568 manufacturing firm-years with disproportionate inventory increases during 1998-2002, among which only 282 observations provide explanations in the MD&A section of 10-K. Further, Sun (2010) classifies the 282 explanations as favorable, neutral and unfavorable and finds that the favorability of the provided explanations has a significant

positive association with a firm's return on assets and sales growth in the subsequent three years. This result suggests that MD&A disclosures on disproportionate inventory increases, if provided, extend financial statement information and help interested users to predict future firm performance.

What drives the amount of forward-looking statements in the MD&A section of 10-K filings is the primary focus of Muslu et al. (2014). The authors create a comprehensive list of future-related keywords and phrases to distinguish forward-looking statements from other statements (such as statements related to historical events, uninformative legal or boilerplate sentences, etc.) in the MD&A section. Using this list, Muslu et al. (2014) extract forward-looking sentences from MD&As with the help of text-parsing software. They test the validity of their methodology by asking human annotators to identify FLS in a set of randomly-selected MD&As, and find that their methodology is both well-specified and powerful. The amount of forward-looking disclosures is measured by the proportion of sentences with forward-looking connotation. The results indicate that firms make more forward-looking MD&A disclosures to improve the information efficiency of stock prices with respect to accounting earnings. The results are stronger for operations-related forward-looking disclosures, disclosures made by loss firms, and disclosures that are made prior to 2000.

There are few papers that analyze the relationship between the level and content of annual disclosures and firm's cost of capital. Botosan (1997) develops a measure of disclosure level that is based on the amount of voluntary disclosure provided in the 1990 annual reports of 122 manufacturing companies. The results of the study indicate that greater disclosure of firms with a low analyst following is associated with a lower cost of equity capital. No significant results are found for firms with a high analyst following. Kothari et al. (2009) extend the study of Botosan (1997) by examining the implications of disclosures by managers, financial analysts, and news reporters (i.e., business press) for firms' information environment. The authors conduct a content analysis of more than 100,000 disclosure reports and find that if disclosures have positive sentiment, the firm's risk (as

measured by the cost of capital, stock return volatility, and analyst forecast dispersion) declines. In contrast, disclosures with negative sentiment lead to significant increases in risk measures. When analyzing the information value of disclosures by source (disclosures by managers, financial analysts, and news reporters), Kothari et al. (2009) find that that negative disclosures from business press lead to increased cost of capital and return volatility, while favorable reports from business press reduce the cost of capital and return volatility. These findings suggest that the tone or sentiment of textual disclosures affect firms' risk and information environment.

A series of corporate failures and accounting scandals that occurred in recent years has provoked an increased attention to the importance of fraud prevention and detection. Research studies have started incorporating both quantitative and qualitative information to build better explanatory and prediction models for fraud and litigation cases. For instance, Goel et al. (2010) examine the verbal content and presentation style of the qualitative portion of the annual reports using natural language processing tools. The authors analyze linguistic features that distinguish fraudulent annual reports from non-fraudulent annual reports and find that adding linguistic features to the analysis improves the overall effectiveness of fraud detection. In related research, Hoberg and Lewis (2013) examine the strategic qualitative disclosure choices of firms involved in potentially fraudulent activity (as indicated by Accounting and Auditing Enforcement Releases (AAER) from the SEC website). The authors conduct textual analysis of annual MD&A disclosures and compare MD&A sections across firms involved in SEC enforcement actions and benchmarks based on industry, size and age. They also look at each firm's own MD&A sections before and after SEC alleged violations. There is a strong evidence that firms involved in alleged fraud anti-herd with industry peers on localized disclosure dimensions and moderately strong evidence that firms herd with industry peers on broader disclosure dimensions. Content analysis helps to identify key vocabulary terms used by firms involved in SEC enforcement actions. The results suggest that firms use complexity to potentially conceal fraudulent actions, and these firms often use uncertain, litigious, and speculative words.

In a related study, Nelson and Pritchard (2007) analyze the use of "cautionary language" in SEC reports and its association with litigation risk. The results indicate that firms facing greater litigation risk use more cautionary statements in their disclosures. This evidence is consistent with strategic disclosure choices by managers to reduce expected litigation costs.

The SEC frequently raises concerns about the informativeness of firms' MD&A disclosures and encourages companies to avoid generic or boilerplate statements.[3] Motivated by the SEC concerns, Brown and Tucker (2011) hypothesize that MD&A is potentially uninformative if it does not change from previous years, especially after significant economic changes at the firm. They develop a measure for narrative disclosure stickiness (MD&A modification score) and find that firms with larger economic changes modify their MD&A more often than those with smaller economic changes. The magnitude of stock price responses to 10-K filings is positively associated with the MD&A modification score suggesting that MD&A has informative to the market content. Surprisingly, MD&A modification scores have declined in recent years, while MD&A disclosures have become longer.

Companies planning to go public file the form S-1 with the SEC. The S-1 is one of the main documents used by investors to research a company in the initial public offering (IPO) process. Loughran and McDonald (2013) examine the tone of the S-1, in terms of its definitiveness in characterizing the firm's business strategy and operations and how it affect investors' ability to value the IPO. To measure the definitiveness of the tone, the authors use Loughran and McDonald (2011)'s uncertain, modal weak, and negative word lists. The results indicate that IPOs with high levels of uncertain text have higher first day returns, absolute offer price revisions, and subsequent volatility. These results are consistent with the theoretical models of uncertainty, book-building, and prospect theory.

---

[3]For instance, the SEC states: "Because matters do not generally remain static from period to period, we would expect [...] change over time to remain current. [...] boilerplate disclaimers and other generic language generally are not helpful in providing useful information or achieving balance, and would detract from the purpose" (SEC Release Nos. 33-8350; 34-48960; FR-72).

2.3 Textual Analysis of Earnings Announcements and Earnings Conference Calls

Until recently, the quarterly earnings announcement and accompanying conference call were not highly regulated activities. In March 2003, however, right after the passage of the Sarbanes-Oxley Act, the earnings release became a part of the reporting requirements under the Exchange Act for the first time. Companies now supply Form 8-K with SEC with press release as exhibit after press release is issued and prior to the conference call. The conference call is typically held within 48 hours, permitting investors to listen to the call. These regulatory changes have increased interest by researchers in information provided in earnings releases and earnings conference calls (Bushee et al., 2004). Researchers have started examining the information content of earnings releases and calls and their implications for the market. For example, Demers and Vega (2008) examine the verbal information contained in texts of management's quarterly earnings press releases. Using a textual-analysis program, Diction, the authors extract various dimensions of managerial net optimism expressed in 21,580 firm-quarter corporate earnings announcements over the period from 1998 to 2006. The results indicate that unexpected net optimism in managers' language affects abnormal returns around announcement periods and predicts post earnings announcement drift. In addition, the authors show that it takes longer for the market to understand the implications of qualitative information than those of quantitative information. They also find that the market reaction varies by firm size, turnover, media and analyst following, and the extent to which the standard accounting model captures the underlying economics of the firm.

Whether the linguistic style of managers is informative about future performance is the focus of Davis et al. (2006). Davis et al. (2006) analyze whether managers use optimistic or pessimistic language in earnings press releases. To analyze the tone of quarterly earnings press releases, the authors employ the linguistic program Diction. Using a sample of 23,400 earnings press releases published on the PR Newswire between 1998 and 2003, they find

a significant positive (negative) association between levels of optimistic (pessimistic) tone in earnings press releases and future return on assets. In addition, they discover a strong market reaction to managers' tone in earnings press releases in a short window around the earnings announcement date. Henry (2006) also examines whether verbal components of firms' earnings press releases improve the prediction of market response to earnings announcements. The verbal content of earnings press releases is captured by elementary computer-based content analysis (tone, length, numerical intensity, and linguistic complexity). Using a sample of 441 companies in 2002, Henry (2006) shows that the inclusion of a broad range of numerical financial variables does not enhance predictive accuracy of market returns, whereas inclusion of verbal variables does result in greater predictive accuracy.

Given that spoken language of earnings conference calls may contain useful information, above and beyond preceding earnings announcements, Price et al. (2012) examine the tone of earnings conference calls and the corresponding market reaction. The results indicate that the linguistic tone of conference call is a significant predictor of abnormal returns and trading volume. Moreover, conference call tone dominates earnings surprises over the 60 trading days following the call. The question and answer portion of the call has incremental explanatory power for the post-earnings announcement drift. In a related study, Engelberg (2008) studies the differential effect of information processing costs in the context of earnings announcements and the post-earnings announcement drift (PEAD). Using a large number of earnings announcements in the Dow Jones index for the period 1999 to 2005, Engelberg (2008) finds that the textual information contributes uniquely to the PEAD phenomenon. The author also shows that certain combinations of words (mainly words related to sales, profits, and income) and future forecasts help to predict future performance and returns. Furthermore, the results indicate that the more confusing the textual information is, the more slowly it is reflected in stock prices.

There are some papers that link textual characteristics of earnings announcements to litigation or litigation risk. Rogers et al. (2011) analyzes the relation between disclosure tone and shareholder litigation. Using a sample of 165 lawsuits filed between 2003 and

2008, the authors find that sued firms use substantially more optimistic language in their earnings announcements than non-sued firms. This difference in tone of disclosures is consistent with the notion that managers issue too optimistic disclosures during periods of losses. In addition, Rogers et al. (2011) examine the combined effect of optimistic language and insider trading. The results indicate that the interaction between optimism and abnormal insider selling is associated with an increased probability of being sued. However, there is no evidence that insider selling (analyzed alone) exposes the firm to increased litigation risk.

Some studies examine the use of deceptive language during the conference calls. Larcker and Zakolyukina (2012) perform an exploratory textual analysis of managers' linguistic statements during conference calls to predict whether the financial reports discussed in the calls are later restated. They find that subsequent restatements are more likely when managers use more references to general knowledge, fewer non-extreme positive emotion words, and fewer third-person plural pronouns. In additional tests, Larcker and Zakolyukina (2012) find that restatements are also more likely when CFOs use more words overall, more negation words, and more swear words, and when CEOs use more extreme positive words, more tentative words, more certainty words, and more hesitations.

Some studies analyze the association between linguistic style of conference calls and insider trading activities. The idea behind these analyses is that managers may benefit from their insider knowledge and manipulate firm-related disclosures to execute their own successful trades. Brockman et al. (2012) find that the positive tone of conference calls predicts insider selling, whereas the negative tone of conference calls predicts insider buying. The inverse tone-trading strategy is stronger for CEOs than for non-CEOs and for small firms than for large firms.

As earnings conference call consists of two sections, introductory remarks by managers and questions and answers (Q&A) section with analysts and investors, several papers explore the information characteristics of both sections. Matsumoto et al. (2011) analyze

more than 10,000 transcripts of conference calls and find that both sections have incremental information value over the earnings press release, but the Q&A section is relatively more informative than the introductory section. This finding suggests that the bigger benefit of conference calls comes from analysts' involvement in the Q&A session. Managers tend to provide more disclosure during the introductory session when firm performance is poor, but relatively more information is released during the Q&A session. Overall, the results are consistent with the notion that active analyst involvement in conference calls increases the information value of the calls, especially when firm performance is poor.

The questions and answers (Q&A) section of earnings conference call involves the use of natural and spontaneous language. Using textual analysis, Chen et al. (2013) examine whether the time of day has an impact on the tone of the Q&A part of earnings conference calls. The results indicate that the tone of the conversations between analysts and managers becomes significantly more negative as the day wears off, holding everything else fixed. More negatively toned conversations are associated with more negative abnormal stock returns during the call and immediately after the call. The authors attribute these results to physical and mental fatigue at later times during the day.

Using earnings press release data, Henry and Leone (2010) evaluate the methods used in accounting and finance literature to measure tone of qualitative disclosures. They argue that different word lists, which are used to identify positive/negative words in disclosures, may lead to a different "tone score" for any given document. Henry and Leone (2010) compare the predictive validity of two commonly-used word lists (GI or Diction) to a unique word list developed for the context of financial disclosures. The results are consistent with Loughran and McDonald (2011) that the context-specific word lists are more powerful than the general word lists.

2.4 Textual Analysis of Financial News and Social Media

The content of financial news or social media about a specific company or the whole stock market may be useful in explaining stock market movements. Several studies in the literature analyze whether the information content in financial and social media is informative about investors' interpretations of stock market performance. For example, Tetlock (2007) uses the General Inquirer (GI), a popular quantitative content analysis program, to analyze daily variations in the Wall Street Journal (WSJ) "Abreast of the Market" columns over the period 1984-1999. The author constructs a measure of media pessimism from the content of the WSJ columns and estimates the association between the constructed measure and the stock market. The results indicate that high levels of media pessimism tend to predict downward pressure on market prices, followed by a reversion to fundamentals. In addition, high or low values of media pessimism are significantly associated with high market trading volume. This finding contradicts the intuition that higher levels of pessimism should lead to lower stock returns.

In a related study, Tetlock et al. (2008) examine the impact of negative words in all Wall Street Journal (WSJ) and Dow Jones News Service (DJNS) stories about individual S&P 500 firms from 1980 to 2004. The results indicate that increases in the number of negative words used in WSJ and DJNS columns about S&P 500 firms relative to prior stories predict larger negative shocks to future earnings. Tetlock et al. (2008) show that a trading strategy that takes a short position in stocks of firms with relatively many negative words in the Dow Jones News Service stories from the previous day, and a long position in stocks with relatively few negatively worded stories, generates significant abnormal returns (excluding transactions costs). In addition, they show that stock market prices gradually incorporate the information in negative words over the next trading day. Overall, these results suggests that the words contained in financial news stories are not redundant and worthless information, but instead capture "hard-to-quantify" aspects of firms' fundamentals.

Financial press claims that Internet stock message boards can move financial markets.

Antweiler and Frank (2004) examine the relation between information from Internet stock message boards and stock returns and volatility. Using the linguistic features of more than 1.5 million messages posted on Yahoo! Finance, the authors find that Internet stock messages help predict market volatility. The effect on stock returns is statistically significant but economically small. In a related study, Chen et al. (2013) analyze the extent to which investor opinions transmitted through social media can predict future stock returns and earnings surprises. Using textual analysis of articles and related commentaries published on a social-media platform for investors, the authors find that the views expressed in both articles and commentaries predict future stock returns and earnings surprises. These findings are consistent with Bollen et al. (2011) study that finds a significant time-series correlation between Twitter feeds and the value of the Dow Jones Industrial Average.

2.5 Summary

Whereas previously researchers were restricted to hand collected and hand coded data, increased and inexpensive computational power has made ubiquitous automated data analysis of a large cross section of firms possible. Researchers have started developing methodologies for fusing quantitative and qualitative data and learning new previously unknown patterns and relationships. Our extensive literature review shows that qualitative information in mandated and voluntary disclosures, financial news, and social media is helpful in explaining current and future firm performance and valuation, above and beyond traditional quantitative factors. Most textual analysis studies in the literature have focused on examining the association between the amount, tone, and readability of qualitative disclosures and future earnings and returns. Table A.1 tabulates disclosure-based papers discussed in this chapter based on the disclosure source, sample size, and textual analysis methodology of these studies. In Chapter 3, we develop a new textual analysis methodology to incorporate MD&A texts into explicit firm-level forecasts of future earnings.

[Table A.1 about here.]

Chapter 3

Earnings Forecasts

3.1 Background

This chapter examines the predictive value of disclosures by assessing the accuracy of forecasts generated with and without the content contained in the Management Discussion and Analysis (MD&A) of 10-K report. After a decade of regulation aimed at improving financial disclosure, it may not be surprising to find that MD&A text is relevant. Moreover, many academic papers have found an association between the amount, tone, and readability of MD&A disclosure and firm fundamentals or market returns.[1] In a sample of firms with disproportionate inventory increases, Sun (2010) finds that the favorability of the explanations provided in the MD&A section has a significant and positive association with a firm's return on assets and sales growth in the subsequent three years. Li (2008) finds an association between readability and current earnings, as well as evidence that after controlling for profitability, more complicated financial reports are associated with lower earnings persistence. While association models rely on theories to suggest causal relationships, they

[1] Some representative papers include Feldman et al. (2009), which finds a significant association between tone change and short window market reactions around the SEC filing date; Li (2010), which shows that firms with optimistic disclosures about future performance have higher future earnings; and Brown and Tucker (2011), which demonstrates that firms with greater changes in MD&A disclosures have greater changes in the following year's performance.

do not imply that the related variables have predictive power (Shmueli, 2010).[2] To our knowledge, this study is the first to develop techniques to incorporate text into explicit firm-level future earnings forecasts and assess the predictive value of textual information. Traditional earnings forecasting models rely only on quantitative financial information, ignoring the fact that most information that flows out of American corporations is qualitative (Gangolly and Wu, 2000). The rapid development of text analysis tools and the availability of 10-K reports in electronic form allows us to analyze high volumes of the qualitative data, provided by firms and regulated by the SEC. Comparing the accuracy of text-enhanced and quantitative models, we assess the relevance of text and provide the launching point for theories of complex relationships between words and future performance that might be hard to hypothesize ex ante.[3]

We identify the words in MD&A with financial sentiment and use feature selection techniques to select a smaller subset of words with the greatest predictive value in forecasting earnings. Although feature selection drastically reduces the number of predictive words, we still have a large variable space, causing problems of dimensionality and potential multicollinearity among selected words. To overcome these issues, we employ several regularized least square methods, including ridge regression, kernel ridge regression (KRR), and lasso regression. These methods that can easily deal with a large number of predictor variables, multicollinearity, and capture both non-linear and linear dependencies among variables.

We say that MD&A is informative if its qualitative content improves forecast accuracy

[2]For example, in a study of Netflix recommendations, researchers found that attributes of the movie, like actors and genre, were significantly associated with an individual's ratings, but predictions of the ratings were less accurate when those movie attributes were included as predictors (Bell et al., 2008).

[3]Shmueli (2010) provides an excellent description of explanatory and predictive modeling.

over models based solely on quantitative information.[4] We predict one-year-ahead ROE using (i) current earnings and its components and (ii) current earnings, its components and and text contained in the MD&A section. We find that models utilizing MD&A disclosures are significantly more accurate (statistically and economically) than models that use only quantitative information.

In addition to providing an approach for incorporating text into forecasts, we look at different firm characteristics to determine those firms which text is the most informative. MD&A guidance calls for firms to provide information about the quality and potential variability of earnings and cash flows so that investors can determine whether the past is indicative of the future. If a firm is relatively risk free, its financial indicators will be strong, and therefore additional text will do little to improve on quantitative forecasts. On the other hand, if a firm is risky, adding context to the financial statements may result in more accurate predictions. We find that firms with large changes in future (past) performance have more (less) informative MD&A sections and when future changes are negative, disclosures are relatively more informative than when future changes are positive.[5] Consistent with providing information about uncertainties that will affect operating performance and changes to liquidity, we find MD&A to be more informative for firms closer to distress and those with higher accruals. Finally, the SEC suggests that firms with more complexity have a greater need to focus their MD&A on material information. We find mixed results on MD&A informativeness and complexity. Our findings offer regulators some evidence that MD&A disclosures are (at least partially) tailored to firms' particular facts and circumstances.

---

[4]It is not obvious that MD&A disclosures will improve the accuracy of earnings forecasts because of their low signal-to-noise ratio. The SEC has repeatedly raised concerns regarding generic 'boilerplate' statements and immaterial and/or redundant details in the section. Consistent with these concerns, Pava and Epstein (1993) find that MD&A sections of 25 randomly selected companies mostly describe past performance (i.e., redundant information). Relatedly, Brown and Tucker (2011) find evidence of increasing MD&A length and decreasing MD&A modifications.

[5]The finding in Sun (2010), that firms with disproportionate inventory increases have significant and positive association between disclosure and return on assets and sales growth indicates disclosure relevance for firms with significant account changes.

Firms are expected to improve the clarity and understandability of their MD&A by using language that is less convoluted.[6] We find that text-enhanced forecasting models are more accurate when MD&A sections are shorter, as measured by the document length. Given that a machine, rather than a human, is "reading" the section, longer text should add no additional processing cost. The finding that concise disclosures have greater predictive value suggests that style captures something more general about the quality of disclosures, which cannot be attributed to information overload or difficulty in processing data.

Li (2010) finds no evidence that MD&A informativeness has increased following recent regulatory reforms. We take a different approach to evaluating the outcome of recent regulatory reforms. Specifically, we calculate whether the gains in forecast accuracy from incorporating narrative disclosure are greater following the regulatory changes. We split our forecast sample into two sub-periods, where the pre-regulation period is 1999-2002 and the post-regulation period is 2003-2010, and analyze the difference of forecast improvements. We find that MD&A information improves forecast accuracy significantly more in the post regulation period than in the pre-regulation period. This suggests that firms may indeed be providing more helpful information, insofar as it can be incorporated into time-series models to lead to better estimates of future performance. There are many differences between the approaches undertaken in Li (2010) and our study that can cause the divergent findings. We use the entire MD&A section (rather than exclusively forward-looking sentences) and we measure informativeness by relative improvements in forecast accuracy rather than difference in language tone.[7]

Our sample contains one of the largest, unanticipated, economy-wide shocks in recent history (2007-2009 financial crisis). Ideally, MD&A would eliminate or reduce earnings uncertainty, but it is only possible if managers are able to predict forthcoming events. We

[6]Motivated by the SEC's instructions that disclosures should be presented in a clear and straightforward manner and avoid duplications that overwhelm readers, several papers examine the cross-sectional differences in readability (see for example Li (2008) You and Zhang (2009), and Miller (2010)).

[7]Indeed, we find little difference across the periods in MD&A tone (i.e., level of positivity and negativity in language), which is consistent with Li (2010) findings.

find that text is less informative overall between 2007-2009, and even more so for firms with low cash positions or firms in the consumer discretionary sector (i.e., those firms most affected by the credit crisis). This finding suggests that even if firms endeavor to provide MD&A with predictive value, they may fail to do so when future earnings are so uncertain because of economic shocks.

Market participants have various sources of information. Many academic papers proxy for earnings expectations with analysts' consensus forecasts stemming from (a) the belief that analysts use multiple sources to generate forecasts and (b) empirical evidence that analysts' forecasts are superior to quantitative time series models (Fried and Givoly 1982; Brown et al. 1987; O'brien 1988).[8] This leaves us with two important questions: Do text-enhanced models models work better than quant-only models for firms that are not followed by analysts? Do analysts genuinely have an advantage at forecasting? We find that text-enhanced models are superior to quantitative models for firms with no analyst following, highlighting the value of incorporating text when alternative proxies for earnings expectations are not available. While analysts appear to be better than text-enhanced models in very competitive forecasting environments or for large firms with many sources of additional information (e.g., press coverage, interim releases), text-enhanced models are superior to analysts in less competitive environments and/or for smaller firms, casting doubt on analysts' superiority over more sophisticated forecasting models.

Our study contributes to the literature in several ways. First, we provide a methodology for improving earnings forecasts using textual disclosures. We avoid losing information through extreme aggregation or human judgment. We can combine our approach with the theories that have emerged from earlier explanatory studies to quantify the effects in those association studies.[9] Second, our analysis identifies complex patterns and relationships (among disclosures) that are too difficult to hypothesize; words which may seem positive

[8]For an excellent summary of the analysts' forecasting superiority literature, see Bradshaw et al. (2012).

[9]For example, our techniques could be used to forecast individual rates of mean reversion using a text matrix based on words describing competition (see Li et al. (2012)).

in a general sense, may be systematically used to sugarcoat something negative. Second, we provide information about the usefulness of disclosures across time periods, firm characteristics and financial performance. While the SEC requires disclosures of all registrants, the disclosure rules may not lead to equally meaningful disclosures over different economic cycles and across firms. We effectively provide the SEC with evidence on whether improvements have taken place where we need them most, or whether the improvements only serve to enhance an already good information environment. We also supply some additional evidence on the importance of using text for firms that are not followed by analysts. Finally, we provide new evidence on the superiority of analysts' consensus forecasts over forecasts generated by text-enhanced time-series models.

The remainder of the chapter is organized as follows. Section 3.2 develops our hypotheses. Section 4.1 discusses the data collection process and variables measurement. Section 4 presents methodologies for earnings forecasting. Section 5 reports the results. Section 6 concludes the chapter.

3.2 Research Questions and Hypotheses

This section discusses main research questions and hypotheses. In Section 3.2.1, we develop two sets of earnings forecasting models: (1) traditional forecasting models that incorporate information from financial statements only (i.e., quantitative benchmarks), and (2) text-enhanced models that in addition to traditional predictors incorporate also text from the MD&A section of 10-K reports filed with the SEC (i.e., alternatives). We hypothesize that if textual information in the MD&A section is irrelevant to earnings forecasting, then there will be no gains in accuracy after adding text.

3.2.1 Text as a Predictor of Future Earnings

To understand how narrative disclosure can improve forecasts of future earnings, we begin with traditional time-series models of annual earnings. These models rely purely on

quantitative (Q) accounting-based variables. Gerakos and Gramacy (2013) show that using just one predictor, e.g., lagged earnings, is often superior to using larger predictor sets that include up to 24 variables. Given that simple models of annual earnings perform at least as well as more complex models, we use the following four (simple and popular) earnings forecasting models:

$$\textbf{Model 1Q}: \quad ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1},$$

$$\textbf{Model 2Q}: \quad ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1},$$

$$\textbf{Model 3Q}: \quad ROE_{t+1} = \beta_0 + \beta_1 CFO_t + \beta_2 ACCR_t + e_{t+1},$$

$$\textbf{Model 4Q}: \quad ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \beta_2 SDD_t + \beta_3 SALES_t + \beta_4 SDD \text{ x } SALES_t + e_{t+1}.$$

where ROE_t is the current return on equity, measured as net income before extraordinary items divided by average book value of owners' equity; $OPINC_t$ is current operating income after depreciation, net of interest expense, special items, and minority interest divided by average book value of owners' equity; $NOPINC_t$ is current non-operating income, net of income taxes, divided by average book value of owners' equity; CFO_t is the current cash flow from operating activities, net of cash flows from extraordinary items and discontinued operations, divided by average book value of owners' equity; $ACCR_t$ is the current accruals component of earnings, measured as the difference between earnings and cash flows, divided by average book value of owners' equity; SDD_t is the sales decrease dummy that takes the value 1 if sales revenue decreases from prior year, and 0 otherwise;[10] $SALES_t$ is the net sales revenue divided by average book value of owners' equity.

Model 1Q uses historical return on equity to predict one-year-ahead return on equity. Model 2Q disaggregates income into its operating and non-operating components scaled by average book value of equity, thus allowing the coefficients to vary based on the relative permanence of each. Model 3Q decomposes income into its accrual and cash flow components and uses these to predict one-year-ahead return on equity. Finally, Model 4Q

[10]Following Banker and Chen (2006), we estimate the sales decrease indicator SDD_t with a logit regression.

identifies cost variability with sales changes (in particular, cost stickiness when sales decline) and uses current sales and indicator of sales decreases in addition to current ROE to predict future ROE (see Banker and Chen (2006) for more details).

Inclusion of textual information into large sample statistical models was difficult (or impossible) until fairly recently. Advances in computing technology and access to electronic filings (on SEC's EDGAR repository) has made it possible to incorporate narrative information into forecasting models for a large sample of firms. However, if early research suggests that simple quantitative models are at least as good as complex ones, why should we include text in forecasting models simply because we *can*?

Understanding the relevance of text in forecasting is important for several reasons. Given increasing volumes of textual data, it is crucial to understand whether textual data is purely boilerplate or it contains incremental information that we can use to complement relevant numerical data to learn previously unknown patterns and relationships. As discussed in Chapter 2, there many studies that find a significant association between MD&A characteristics and firm performance. However, in-sample explanatory power is not equivalent to out-of-sample predictive power. While explanatory power shows the strength of an underlying causal relationship, it does not imply its predictive power. Therefore, predictive models could serve as a "reality check" to the relevance of management discussions in 10-K reports. Shmueli (2010) gives an excellent overview of the differences between explanatory and predictive power.

Without testing the outputs from predictive models, we do not know the possible gains to accuracy from including qualitative disclosures. The SEC has emphasized the importance of and imposed significant requirements on narrative disclosures to enhance their usefulness. However, if disclosures do not improve predictions, the MD&A may not be relevant in helping investors to predict future performance. The very nature of MD&A and

its safe-harbor provisions may make it relatively costless for managers to provide misleading information.[11] Incorporating text into forecast models enables us to address whether the protections of MD&A effectively allow for opportunism that diminishes (rather than enhances) the value of narrative disclosures.[12] Last, a long literature suggests that analysts are more accurate than time-series models and attributes analysts' superiority, in part, to a contemporaneous advantage (i.e., more information). Therefore, it seems possible that by incorporating text, we can enhance accuracy of time-series forecasts by capitalizing on the contemporaneous advantage while avoid the conflicts of interests that lead to biased low quality forecasts. Moreover, our cross-sectional analysis allows us to generate earnings forecasts for firms that have no analyst coverage or long time series of earnings realizations.

We enhance the quantitative models described above with text in two ways. First, we create variables for the six word lists developed by Loughran and McDonald (2011) (i.e., positive, negative, litigious, uncertain, modal weak and modal strong) and use these variables as predictors in our models (*CategMatr*). Loughran and McDonald (2011) analyze a large number of SEC reports to develop those words lists and show that some of their word lists are significantly associated with market reactions and unexpected earnings. They also emphasize the importance of using domain specific word lists for textual analysis. Each of the six variables in *CategMatr* is set equal to the weighted sum of frequencies of words in that category in the MD&A (i.e., sum of frequencies of words in the category divided by total words or other term weighting factor, see Section 4.1.2 for more details) in a given

[11] As long as there is no explicit evidence that the manager acted without good faith or a reasonable basis, forward looking information that turns out to be inaccurate is protected. Consistent with an argument that safe-harbor provisions may make it relatively costless for managers to provide misleading information, Li (2008) provides evidence that MD&A section can be used as an 'obfuscation tool' to mask poor performance.

[12] Rogers et al. (2011) find that sued firms have more optimistic language in their earnings announcements than non-sued firms, and there is an increase in the likelihood of being sued if the firm uses optimistic language while its insiders are selling their shares.

firm-year.

$$
\begin{aligned}
\textbf{Model 1C}: ROE_{t+1} &= \beta_0 + \beta_1 ROE_t + \alpha CategMatr_t + e_{t+1}, \\
\textbf{Model 2C}: ROE_{t+1} &= \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha CategMatr_t + e_{t+1}, \\
\textbf{Model 3C}: ROE_{t+1} &= \beta_0 + \beta_1 CFO_t + \beta_2 ACCR_t + \alpha CategMatr_t + e_{t+1}, \\
\textbf{Model 4C}: ROE_{t+1} &= \beta_0 + \beta_1 ROE_t + \beta_2 SDD_t + \beta_3 SALES_t + \beta_4 SDD \text{ x } SALES_t \\
&+ \alpha CategMatr_t + e_{t+1}.
\end{aligned}
$$

While this approach simplifies the estimation process dramatically, it also imposes several significant assumptions. First, six word categories are subjective and it may be the case that such categorization is not informative about future earnings. Second, predetermined text categories implicitly assume that all words in the category are equally informative. For example, when creating a category of positive words, we simply add normalized frequencies of positive words, ignoring a possibility that some words are more relevant to future earnings than others. Words that measure positive, negative, uncertain, etc. tone of a document can vary significantly in their relevance, strength and frequency (e.g., catastrophic vs. reluctant, exceptional vs. effective). Therefore, a more detailed analysis of texts is desirable.

The second approach we propose does not aggregate words into (pre-determined) word categories, but instead lets the model determine the predictive power and "meaning" of individual words. Each word is counted and normalized separately, and a statistical model is used to determine the set words that can be used as predictors in a given period (see Section 4).[13] This allows specific words to have different relevance and predictive power

[13]For consistency of results, we also use a feature selection algorithm for six weighted word categories. All the results are qualitatively the same. We do not perform a feature selection on quantitative financial statements variables, instead we rely on earnings prediction models developed and tested in numerous prior studies. Section 4 provides more details on the methodology used in the paper.

across firms and years.

$$
\begin{aligned}
\textbf{Model 1T}: ROE_{t+1} &= \beta_0 + \beta_1 ROE_t + \alpha TextMatr_t + e_{t+1}, \\
\textbf{Model 2T}: ROE_{t+1} &= \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha TextMatr_t + e_{t+1}, \\
\textbf{Model 3T}: ROE_{t+1} &= \beta_0 + \beta_1 CFO_t + \beta_2 ACCR_t + \alpha TextMatr_t + e_{t+1}, \\
\textbf{Model 4T}: ROE_{t+1} &= \beta_0 + \beta_1 ROE_t + \beta_2 SDD_t + \beta_3 SALES_t + \beta_4 SDD \text{ x } SALES_{\ t} \\
&+ \alpha TextMatr_t + e_{t+1}.
\end{aligned}
$$

$TextMatr_t$ is the $m \times n$ matrix of normalized (by the total number of MD&A words or other normalization variable, see Section 4.1.2) textual features in current period t, where n and m are the number of selected words and observations, respectively. The coefficient vector for textual features α has dimension $n \times 1$, where n is the number of textual variables that are selected for *ROE* prediction.

Models 1Q-4Q are parsimonious models that include only quantitative information from financial statements. Models 1C-4C and 1T-4T are larger models that in addition to quantitative information include also textual information from MD&A (i.e., word categories for C-models, and individual words for T-models). Larger (C or T) nest Models 1Q-4Q. In other words, each model C or T can be reduced to model Q if parameters on the corresponding text variables are set to 0. For example, Model 2C can be reduced to Model 2Q if parameter α on *CategMatr* is set to 0.

Under the null hypothesis we state that additional textual variables in Models 1C-4C and 1T-4T do not help prediction and the forecast errors of Models 1Q-4Q should be smaller than that of corresponding larger model that incorporates text. Intuitively, each parsimonious model (in our case, Models 1Q-4Q) gains efficiency by setting to zero parameters that are zero in population, while the alternative (in our case, Models 1C-4C and 1T-4T) introduces noise into the forecasting process that will inflate its forecasts errors in finite samples (see Clark and West (2007); Clark and McCracken (2013) for more details). Stated formally, we hypothesize that:

Hypothesis 1.

(a) Quantitative models (Q) are less accurate than models enhanced with text categories (C).

(b) Quantitative models (Q) are less accurate than models enhanced with selected text (T).

Tests of H1 allow us to see whether adding textual information from the MD&A section results in superior forecasts of earnings and also which textual approach (aggregation of textual information using predefined categories or more detailed word-by-word analysis) works better.

We limit our attention to models that are based solely on financial statements data or financial statements data and MD&A text. Although we do not rely on a wider array of quantitative and qualitative information in this paper, our methodology can be used to combine a broader range of textual information as well as economy- , industry-, or firm-specific numerical information.

3.2.2 Determinants of MD&A informativeness

Quantitative models may not perform well in forecasting earnings if a company faces high underlying uncertainty. For example, when the company experiences or expects large changes in performance, has earnings with a high accrual component and thus low earnings persistence, or is in the state of financial distress. Moreover, the SEC's guidance seems to impose a greater duty to disclose on some firms than others and in turn suggests the benefits to using text-enhanced models may be greater for some firms than others. For example, the SEC states that MD&A should: 1) 'allow readers to understand the effects of material changes and events and known material trends and uncertainties [...], and include uncertainties or trends that are reasonably likely to result in the registrant's liquidity increasing or decreasing in any material way'; 2) 'provide information about the quality of, and potential variability of, a company's earnings and cash flows; and 3) 'be presented in clear and understandable language', regardless of firm complexity. If the company is relatively

risk free, its financial indicators will be strong, and therefore additional text will do little to improve on quantitative forecasts. Alternatively, if the company is very risky, adding context to the financial statements may result in better predictions.

These objectives guide our predictions about the relation between firm characteristics and disclosure informativeness. There is the implicit presumption that the firm can provide information that improves understanding, even when situations are uncertain or complex. To control for differences in inherent predictability, we use the S&P quality ranking, which attempts to measure the stability of earnings. If this does not provide an adequate control, a finding different from our predictions can either be consistent with management choosing not to provide or with management being unable to provide relevant information.

We explore several factors to determine whether the predictive ability of the MD&A section varies by firm characteristics. This analysis allows us to see whether MD&A content improves an already good information environment or helps explain uncertainties and risks. Firm characteristics we examine are the following:

Current Performance. MD&A should describe the quality of current financial performance for the purpose of allowing users to assess variability of the earnings stream. Earnings are mean reverting, yet favorable performance is more persistent than negative performance. Therefore, we expect MD&A to be more informative when changes are large and positive.

Future Performance. Firms with large expected changes in earnings are required to provide more informative MD&A disclosures. We expect a positive relation between MD&A informativeness and absolute changes in future earnings. Given that investors react more strongly to negative news than to positive, we also analyze the direction of earnings changes (increase or decrease). If managers expect future earnings to decrease, they may provide more informative MD&A statements to tone down the negative news.

Accruals. Earnings with a high accrual component are less persistent than earnings with a high cash flow component, due to the estimates involved in the accrual component (Sloan, 1996; Richardson et al., 2005; Dechow and Ge, 2006). If investors understand

the implications of accruals for future earnings, then managers have a greater incentive to support the high accruals with an explanation in their MD&A. Alternatively, if investors do not differentiate between accruals and cash flows, we would expect no relation or a negative relation between text informativeness and accruals.

Firm Size. Larger firms face greater SEC and investor scrutiny; their higher political and/or legal costs are likely to induce them to provide more informative disclosures. We expect a positive relation between MD&A informativeness and firm size.

Market-to-Book Ratio. The Market-to-Book (MTB) ratio is widely used as a measure of performance, future growth or efficiency. To the extent that growth firms operate in more uncertain environments, we expect a positive relation between MD&A informativeness and MTB.

Altman's Z-Score. The Z-score, initially developed as a bankruptcy prediction score, is used frequently to measure financial health, with high Z-scores indicating healthy firms and low Z-scores indicating distressed firms. More distressed firms face greater uncertainty and risks. Therefore, we expect a negative relation between MD&A informativeness and Z-score.

Intangibles. The financial statements of intangibles intensive firms tend to be less useful because accounting does such a poor job recognizing the future benefits of intangible investments (Lev, 2001). Firms with high levels of intangible assets need to provide relatively more context for interpreting their financial statements. This suggests a positive relation between MD&A informativeness and reported intangibles.

Audit Quality. Prior research suggests that auditors of top audit companies provide higher quality audits, ensuring the reliability of financial statements (Khurana and Raman, 2004; Francis and Yu, 2009). Although MD&A section is not audited, auditors may examine it to help the firm ensure that its presentation is consistent with SEC rules, the amounts cited within the section are accurate and the underlying data provide a reasonable basis for the forward-looking information contained in the section. We predict a positive relation between MD&A informativeness and audit quality.

Firm Complexity. Because the SEC requires clear and understandable language regardless of firm complexity, firms with complex operations must provide relatively more description than simple firms to enhance existing financial statement disclosures. We predict a positive relation between MD&A informativeness and complexity.

MD&A Transparency (Readability and Report Length). MD&A is considered to be more (less) informative if it is written in a clear, straightforward (unclear, obtuse) way. Li (2008) finds that readability and report length proxy for disclosure transparency (i.e., longer and less readable disclosures are harder to analyze). Although a computer should not have difficulty interpreting longer disclosures or disclosures with more sophisticated language, the decision to write the MD&A in an unclear or convoluted way may be emblematic of an overall managerial preference for less helpful disclosure. Therefore, we expect a negative relation between MD&A informativeness and both readability and report length.

We summarize the predictions on the relation between firm characteristics and text informativeness below.

Hypothesis 2.

> *MD&A disclosures are more informative for firms with large changes in current and future performance, positive changes in current and future performance, high accruals, high market capitalization, high market-to-book ratios, low Z-scores, large amounts of intangible assets, high audit quality, high complexity, and high disclosure transparency.*

3.2.3 Changes in MD&A Informativeness

Regulation relating to MD&A was initially imposed in 1980 but largely untouched until several large corporate failures and accounting scandals.[14] In December 2001, the SEC returned its attention to MD&A, proposing that firms include a section describing the accounting policies that required difficult and subjective judgments and were important to

[14] Bryan (1997) provides some evidence on the success of early regulatory actions regarding MD&A information content. Using a sample of 250 firms in 1990, he finds that MD&A disclosures are significantly associated with one-year-ahead changes in sales and earnings per share.

the portrayal of the company's financial condition and results. In the following year, the SEC addressed the need for additional disclosures related to capital resources and liquidity including off-balance sheet arrangements, related-party transactions, and non-exchange traded contracts accounted for at fair value. Also in 2002, Congress passed the Sarbanes-Oxley Act (SOX) which contains a number of provisions that enhance the quality of financial reporting. SOX requires that both the CEO and the CFO certify the fairness of financial reports, and imposes significant criminal civil penalties for corporate misconduct. SOX also significantly expands the responsibilities of audit committees. Bainbridge (2007) emphasizes that the Sarbanes-Oxley Act intensified the MD&A disclosure requirements even more. Finally, in December of 2003, the SEC issued an interpretation intended to encourage a reduction in boilerplate language and elicit more meaningful disclosures in MD&A. It placed an emphasis on the analysis of known trends, events and uncertainties.[15] As of December 31, 2003, the expectations of managers to communicate with investors in their MD&A section about their firms' current and future performance had been heightened. We compare the accuracy differences between T-models and Q-models before and after the recent reforms to determine whether the narrative disclosure contained in MD&A is more relevant (i.e., has greater predictive value) in the post-regulation period.

Our post-regulation period also covers one of the largest, unanticipated shocks to credit markets in recent history.[16] Although the SEC calls for information allowing investors to ascertain the likelihood that historical financial data are (or are not) indicative of future results, it may be impossible to receive such information when managers themselves do not have it. Therefore, we predict that MD&A informativeness decreases during the crisis period. Because companies and industries were affected differently by the crisis, we test the relation between cross-sectional differences in firms and MD&A informativeness. Under normal circumstances, firms should provide more disclosure when their past performance

[15] See: SEC Release Nos. 4936, 33-8040, 33-8056, 33-8350.

[16] According to the U.S. National Bureau of Economic Research, the crisis began in December 2007 and ended in June 2009.

is not a good predictor of future performance; however, during the 2007-2009 crisis, these changes were mostly unknown and thus their MD&A could not adequately foreshadow the subsequent earnings changes. We expect firms with larger cash positions to be less affected by a crisis (due to their greater flexibility and lesser need to access the financial markets), and therefore we hypothesize that they experience a smaller decline in MD&A informativeness. We also expect that in a financial downturn, consumers tighten their discretionary and luxury spending, and therefore firms in the consumer discretionary sector should be more affected by the crisis.[17]

We summarize the predictions on the relation between time periods and text informativeness below.

Hypothesis 3.

(a) Accuracy improvements from using text-enhanced over quant-only models is greater in the post-regulation period than in the pre-regulation period.

(b) Accuracy improvements from using text-enhanced over quant-only models is lower in the crisis period, particularly for firms in the consumer discretionary sector and firms with low cash positions.

3.2.4 Analysts' Advantage

It is common practice to proxy for earnings expectations using analysts' consensus forecasts. For a large sample of firms, consensus forecasts are readily available and analysts have been shown to be superior to time-series models. While Bradshaw et al. (2012) discuss some of the limitations of these earlier studies, they find that analysts are more accurate than a random walk in short-horizon forecasts, for both large and small firms. Our model incorporates both quantitative and qualitative information into forecast, so a natural question is whether analysts are still better than statistical models. Before answering that

[17]Companies in the consumer discretionary sector include retailers, media companies, consumer services companies, consumer durables and apparel companies, and automobiles and components companies. Consumer staples sector, in contrast, sells products (or provides services) that customers cannot live without or are relatively inexpensive and thus demand is unaffected by economic conditions.

question, however, we note that many firms in our sample are not followed by analysts (around 45%). We predict that text-enhanced models are superior to quant-only models for firms without analyst following, validating the benefits of using text-enhancements when there are no alternative market proxies. For the subset of firms followed by analysts, we conjecture that text enhanced models put statistical forecasting on a more equal footing. We expect that the more competitive the analyst environment is, the more analysts will work to perfect their forecasts (e.g., to make All Star lists, increase their prestige). However, when the competition is limited, they do not necessarily gather more information suggesting that text-enhanced models might capture some of the increased accuracy previously attributed to superior skills. Analysts have access to additional sources of information (e.g., industry reports, press releases, news articles, etc.) than our text-enhanced models. We expect that the amount of additional information is increasing in firm size (overall interest in the firm) and therefore we predict that analysts' forecasting superiority persists for large firms, but disappears for smaller firms.[18] Our predictions are summarized below.

Hypothesis 4.

(a) Text-enhanced models are more accurate than quantitative models for firms without analyst following.

(b) Analysts' superiority over text-enhanced models is increasing in firm size and analyst competition.

Following Frankel and Lee (1998) and Banker and Chen (2006), we calculate analysts' consensus forecasts of ROE comparable to forecasts generated by our text-enhanced time-series models. Specifically, we compute the consensus forecast as the mean of analysts' forecasts of earnings per share reported on I/B/E/S nine months before fiscal year-end, divided by the beginning-of-year average book value of equity per share. In this manner the timing of consensus forecasts roughly coincides with the availability of 10-K reports.

[18] In additional analyses, we also test analysts' forecasting superiority over text-enhanced models for firms with high/low uncertainty, where uncertainty is measured by analyst forecast dispersion. We find that text-enhanced models outperform analysts' consensus forecasts for firms with high analyst disagreement.

The correlation between firm size and analyst following in our is 61%, suggesting that the measures capture (somewhat) different constructs.

Chapter 4

Methodology

4.1 Data collection and Numerical Representation of Texts

4.1.1 Data

Our quantitative models use earnings and the book value of equity. We use income before extraordinary items as our measure of earnings and calculate return on equity (*ROE*) as earnings divided by the average book value of owners' equity. Model 2 requires information on the operating and non-operating components of income, where *OPINC* is current operating income after depreciation, net of interest expense, special items, and minority interest, and *NOPINC* is current non-operating income, net of income taxes, scaled by the average book value of owners' equity. Model 3 requires information cash flow and accrual components of earnings, where *CFO* is the current cash flow from operating activities, net of cash flows from extraordinary items and discontinued operations, divided by average book value of owners' equity, and *ACCR* is the current accrual component of earnings, measured as the difference between earnings and cash flows, divided by average book value of owners' equity. Model 4, in addition to earnings, requires information on sales revenues (*SALES*). These quantitative variables are extracted from COMPUSTAT

for the period 1995-2011, for the entire population of firms. For the subsample of firms with analyst following, we use analyst forecasts of earnings per share reported on I/B/E/S nine months before fiscal year-end, divided by the beginning-of-year average book value of equity per share.

To alleviate concerns about outliers and/or data errors, we exclude firm-year observations (as in Banker and Chen (2006)) that have: (1) missing values on selected variables; (2) negative values of owners' equity; (3) absolute values of *ROE* greater than 1; (4) absolute values of net profit margin greater than 1; and (5) absolute values of percentage change in sales revenue greater than 1; and (6) firms in financial industries (SICs 6000–6999). This results in 45,740 firm-year observations with all the necessary accounting-based quantitative variables.

To combine these observations with textual information, we download 10-K reports using the SEC indices of all filings submitted to the commission. From the 10-Ks, we extract the MD&A section (Item 7 and 7a) in each filing. Because 10-K reports do not have a standardized structure of text, MD&A sections without clear designations are harder to extract. We successfully gather the MD&A section for 91.8% of the total 10-K downloads.[1] Most of the downloaded filings contain HTML tags in addition to the filing text. Following Li (2008, 2010), we delete all the HTML tags, special symbols, stop words, tables and numbers from MD&A documents. These cleaned MD&A files are used to build the qualitative dataset, which we then merge, on CIK and date with COMPUSTAT and then I/B/E/S. Table A.2 summarizes the data population after each processing step. Our final sample consists of 23,976 firm-year observations over the 16-year forecast period 1995-2011.[2]

[Table A.2 about here.]

[1] We do random document comparisons to ensure the accuracy of the extractions.

[2] Many observations were deleted while matching numerical and textual datasets due to the inability to find an appropriate GVKEY for the CIK code specified in the header of each 10-K filing. Chen and Li (2013) report similar sample reductions in their study.

4.1.2 Numerical Representation of Texts

To create textual variables for use in statistical models, we employ the bag-of-words approach (BOW), which identifies each word and counts the number of times it appears in a document. Each MD&A document is represented by the words it contains, with ordering and punctuation ignored. Although this approach is very popular and simple, it results in a partial loss of information. For example, in our analysis, the word 'outstanding' would be coded identically whether it refers to 'the number of shares outstanding' or to 'outstanding sales growth.'

To restrict our analysis to financially relevant words, we use a stemmed version of the Loughran and McDonald (2011) Financial Sentiment Dictionaries (LMFSD) as the primary source for word counts.[3] LMFSD contains 3,532 distinct words that are grouped into six categories: positive, negative, uncertainty, litigious, strong modal and weak modal. The stemmed version of LMFSD contains 1,389 words. Loughran and McDonald (2011) show that some of these word lists are significantly related to the market reaction around the 10-K filing dates, trading volume, unexpected earnings, and subsequent stock return volatility.[4]

Many studies in the automatic text retrieval literature show that weighting counted words properly (often referred to as term weighting) can enhance the effectiveness of an information retrieval system (Salton and Buckley, 1988; Buckley, 1993; Jurafsky and James, 2000). Term weighting improves simple word counts by combining three popular factors that affect the importance of a word in text:

1. *Term Frequency (tf)* - measures the number of times a term occurs within a document;

[3] Stemming is popular in the text retrieval literature method to reduce inflected words to their stem form. Terms with a common stem usually have similar meanings (e.g., abandon, abandoned, abandoning, abandonment, abandonments, abandons), making it more efficient to consider related words as one term, rather than distinct terms.

[4] Loughran and McDonald (2011)'s financial sentiment dictionaries exclude most of the accounting terms frequently used to discuss the strength of a company, its assets, investment and financing strategies, operating activities, etc. (e.g., liabilities, expenses, revenues, depreciation, among many others). Therefore, it is possible that adding an accounting glossary to the list of words with financial sentiment could enhance the textual analysis of linguistic firm-related disclosures.

2. *Inverse Document Frequency (idf)* - downgrades frequently occurring words in a sample of documents, assuming that words that are used in many documents are less informative;

3. *Total Number of Words (tw)* - serves as a normalization for document length.

In our analyses, we use six term weighting rules that account for common words within all MD&A documents, MD&A documents of the same industry, and MD&A documents of the same company.[5] We use $tf_{i,d}$ to denote the number of times a word i appears in a document d; tw_d to denote the total number of words in the document d; a_d to measure the average word count in the document d; N (N_k) [N_c] to denote the number of MD&A documents in the whole sample (in the industry k's sample) [in the company c's sample]; and df_i ($df_{i,k}$) [$df_{i,c}$] to denote the number of (industry k's) [company c's] MD&A documents that contain at least one occurrence of word i. Then, the weighted measure of the i^{th} word occurrence in the d^{th} document is defined as:

(1)
$$w_{i,d} = \frac{tf_{i,d}}{tw_d} \times log(\frac{N}{df_i}) = tf \times idf,$$

where the first term measures the normalized word frequency (i.e., normalized *tf*), while the second term downgrades common words across all MD&A documents (i.e., sample *idf*).

For example, the word 'could', classified as a modal weak word in financial sentiment dictionaries, may be less important if it occurs in most or all MD&A documents from the sample.

(2)
$$w_{i,d} = \frac{1 + log(tf_{i,d})}{1 + log(a_d)} \times log(\frac{N}{df_i}) = log(tf) \times idf,$$

[5]We tabulate the results for the first term-frequency measure. The results for other measures are qualitatively similar.

where the first term decreases the impact of high frequency words with a log transformation and the second term assigns lower weights to more common words in the sample as in (1).

(3)
$$w_{i,d} = \frac{tf_{i,d}}{tw_d} \times log(\frac{N_k}{df_{i,k}}) = tf \times idf_k,$$

where the first term measures the normalized word frequency (i.e., normalized *tf*) as in (1), while the second term downgrades common words across MD&A documents of the same industry (i.e., industry-specific *idf*).

For example, the word 'ideal', classified as a positive word in financial sentiment dictionaries, may be less important if it occurs in most or all MD&A documents of industry k.

(4)
$$w_{i,j} = \frac{tf_{i,j}}{tw_j} \times log(\frac{N_c}{df_{i,c}}) = tf \times idf_c,$$

where the first term measures the normalized word frequency (i.e., normalized *tf*), while the second term downgrades common words across MD&A documents of the same company (i.e., company-specific *idf*).

For example, the word 'claims', classified as a negative word in financial sentiment dictionaries, is less important in the context of the MD&A of an insurance company which typically discusses customer claims.

(5)
$$w_{i,d} = \frac{1 + log(tf_{i,d})}{1 + log(a_d)} \times log(\frac{N_k}{df_{i,k}}) = log(tf) \times idf_k,$$

where the first term is the logged term frequency as in (2) and the second term is the industry-specific idf as in (3).

(6)
$$w_{i,j} = \frac{1 + log(tf_{i,d})}{1 + log(a_d)} \times log(\frac{N_c}{df_{i,c}}) = log(tf) \times idf_c,$$

where the first term is the logged term frequency as in (2) and the second term is the company-specific idf as in (4).

To illustrate how we calculate word weights in each MD&A section, consider the following example based on extracts from two MD&A sections of SunGard Data Systems Inc. (words that bear financial sentiment are in italics):

MD&A$_1$: "We believe that our existing cash resources and cash generated from operations will be *satisfactory* to meet our operating requirements, debt repayments, contingent acquisition payments and ordinary capital spending needs for the foreseeable future. [...] If customers *cancel* or *refuse* to renew their contracts, or if customers reduce the usage levels or asset values under their contracts, there could be a material *adverse* effect on SunGard's business and financial results."

MD&A$_2$: "If customers *cancel* or *refuse* to renew their contracts, or if customers reduce the usage levels under their contracts, there could be a material *adverse* effect on SunGard's operating performance.[...] It is possible that the businesses we have acquired and businesses that we acquire in the future may perform *worse* than expected or prove to be more *difficult* to integrate and manage than expected."

Our calculations are the following:

MD&A$_1$ & **MD&A$_2$**: Total words in $MD\&A_1$: 69. Total words in $MD\&A_2$: 64. Number of documents: 2. Common financial sentiment words in two documents: *cancel, refuse* and *adverse*. The BOW representation of two documents is:

$$\begin{array}{c} \\ MD\&A_1 \\ MD\&A_2 \end{array} \begin{array}{c} \begin{array}{cccccc} \textit{satisfactory} & \textit{cancel} & \textit{refuse} & \textit{adverse} & \textit{worse} & \textit{difficult} \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}$$

Then, using the $tf \times idf$ rule we defined above, we get that the normalized representation of $MD\&A_1$ and $MD\&A_2$ is:

$$
\begin{array}{c} \\ MD\&A_1 \\ MD\&A_2 \end{array}
\begin{array}{c}
\begin{array}{cccccc} \textit{satisfactory} & \textit{cancel} & \textit{refuse} & \textit{adverse} & \textit{worse} & \textit{difficult} \end{array} \\
\begin{pmatrix} 0.010 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.011 & 0.011 \end{pmatrix}
\end{array}
$$

In this example, words *cancel, refuse* and *adverse* will be ignored because they occur in every document, and words *satisfactory, worse* and *difficult* will be considered.

4.2 Feature Selection

In Section 4.1.2, we discussed the numerical representation of texts using the 'bag-of-words' (BOW) approach. While the most notable benefit of BOW is simplicity (i.e., each document is represented by normalized frequencies of words it contains), one of the biggest drawbacks is the dimension of the word-count vector. In the stemmed version of LMFSD, there are 1,389 unique words with financial sentiment and it may be the case that some of the words are just noise or redundant information. To select the most useful words for earnings forecasting, we consider feature selection methods. In Section 4.3, we also discuss a regularized least squares estimation method, called lasso regression, that performs feature selection and estimation simultaneously.

Feature selection methods bring many potential benefits to the forecasting models, including improved prediction performance, lower processing costs, and better data understanding (Guyon and Elisseeff, 2003). There are many feature selection methods that were shown to work well for prediction (for example, principal component analysis, factor analysis, stepwise regression, correlation analysis, etc.). In this study, we do not analyze which feature selection method works best on MD&A texts, but rather emphasize the importance of using such methods, especially in the text mining area, where one has to deal with hundreds of words and word combinations. In our analyses, we use one of the most popular techniques for selecting a subset of relevant features for building robust forecasting models called stepwise regression. Stepwise regression adds and removes variables from a model

based on their statistical significance. The method starts with an initial model (which can be 0, some specific, or all variables from the set) and then checks the explanatory power of two competing models: larger with more variables included and smaller that excludes those variables. At each iteration, stepwise regression calculates the p-value of a variable to determine whether that variable adds any power to the model. The null hypothesis is that the variable has no explanatory power. Therefore, if there is sufficient evidence to reject the null, the variable is added to the model, otherwise the variable is deleted from the model.

To select the most relevant words for earnings prediction, we perform forward stepwise regression feature selection (FSRFS). In FSRFS, there is no order among the candidate independent variables. All variables are evaluated at each step and the best variable is added to the final model. Forward SRFS begins with zero variables in a model, then it checks possible predictor variables one by one, and includes them in the model only if they are statistically significant. Forward stepwise selection stops when all relevant independent variables have been added (Draper and Smith 1966; Zho et al. 2006). In contrast, backward stepwise regression feature selection (BSRFS) begins with all the variables in a model, then it checks possible predictor variables one by one, and deletes them if they are not statistically significant. The goal of both stepwise selection methods is to reduce the computational costs, avoid over-fitting, and build statistical models with the minimal out-of-sample prediction errors.

We apply feature selection to the entire cross-section of firms within an estimation period. Textual features are selected separately using a 4-year rolling window, such that the estimation period (i.e., training period) is $(t-4)$ to $(t-1)$, allowing us to generate out-of-sample forecasts in period t. Figure A.1 summarizes the estimation timeline. Although the feature selection technique reduces the number of predictive words by about 90%, the predictor set (i.e., number of variables) is still large. On average we have 140 distinct words selected for ROE prediction.

[Figure A.1 about here.]

To sum up, stepwise feature selection provides many benefits in forecasting: more parsimonious models, lower processing costs, and better predictors. However, its selected subsets of features can be extremely variable - at each step variables are either added or dropped from the model based on their explanatory power. Small changes in the data or in the initial model can result in very different models being selected (Tibshirani, 1996). In this sense, stepwise models are 'locally optimal', i.e., there is no guarantee that same features would be selected with different inputs. In the next section, we discuss regression based estimation methods that deal with problems of high-dimensionality and multicollinearity in data.

4.3 Estimation Methods

This section reviews popular estimation methods that can be used for earnings forecasting with many predictor variables that are not independent. In Section 4.3.1, we briefly review the Ordinary Least Squares (OLS) method and highlight its drawbacks in dealing with non-linearity and high-dimensionality of data. Further, we introduce the Ridge Regression (RR) method that addresses the problem of multicollinearity by introducing a penalty term in the obejctive function. In Section 4.3.2, we show how Kernel Ridge Regression (KRR) overcomes the drawbacks of multicollinearity and high-dimensionality of data by means of the so-called *kernel trick*. We also discuss main properties of kernel functions that make kernels very popular. Finally, in Section 4.3.3, we introduce the Lasso Regression (LR) method that deals with multicollinearity in data and performs automatic variable selection simultaneously.

4.3.1 Ridge Regression

Consider the usual linear regression model with N predictor variables, $X_1, X_2, ..., X_N$, and the response variable y:

$$y = \alpha + w_1 X_1 + w_2 X_2 + ... + w_N X_N = \alpha + Xw, \tag{4.1}$$

where $y = (y_1, y_2, ...y_L)$ is the vector of companies' *ROE*, α is the vector of ones (i.e., constant term), $w \in \mathbb{R}^N$ is the weight vector, X is the $L \times N$ matrix whose rows are observations and columns are predictor variables, L is the number of observations in the cross section, and N is the total number of predictor variables in the model. The goal is to fit the data and get the coefficient estimates $\hat{\alpha}$ and $\hat{w} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$. When selecting a model estimation technique, typically two aspects are considered: (1) accuracy of prediction on previously unseen data and (2) model interpretation (parsimonious models are always preferred, especially when the number of predictors is large). In Section 4.2, we discussed stepwise feature reduction techniques that can be applied to the data before the actual estimation process takes place. Below we discuss alternative methods that perform feature selection and estimation simultaneously.

Given the large number of predictor variables relative to the number of observation (i.e, L vs. N), estimating Equation (4.1) may pose a challenge. A popular way of estimating a linear model is the ordinary least squares (OLS) method which minimizes the residual sum of squares to get the estimates of α and w:

$$\begin{aligned} L_{OLS} &= (y - \alpha - w_1 X_1 - ... - w_N X_N)(y - \alpha - w_1 X_1 - ... - w_N X_N) \\ &= ||y - \alpha - Xw||_2^2. \end{aligned} \tag{4.2}$$

The unique solution of OLS problem, if it exists, is:

$$w = (X'X)^{-1} X'y, \tag{4.3}$$

OLS often does a poor job in both prediction and interpretation (Tibshirani, 1996; Zou and Hastie, 2005). With respect to prediction, the OLS coefficient estimates usually have low bias but large variance and prediction performance could be improved by setting some of the coefficients to 0. In this manner, the variance of predicted values is reduced at cost of increasing the bias. With respect to interpretation, it can be challenging to interpret the OLS output when the number of predictors is large. Finally, coefficient estimates for OLS regression models rely on the independence of the predictors. When predictors are correlated and have an approximate linear dependence, the matrix $(X'X)^{-1}$ becomes close to singular (i.e., $X'X$ is not invertible). This makes the OLS estimates highly sensitive to random errors and outliers in the observed response y, producing large variance estimates with limited predictive power.[6]

Ridge regression (Hoerl and Kennard, 1988) effectively addresses the problems with OLS by minimizing the residual sum of squares, subject to a bound on the magnitude of the coefficients (L_2-penalty on the coefficients). Ridge regression achieves its better prediction performance through a bias-variance trade-off. Ridge regression shrinks the coefficients and hence is more stable, but it always keeps all the predictors in the model and to get a parsimonious model one has to use a feature selection algorithm separately (before the estimation). In contrast to ordinary least squares, ridge regression minimizes the following loss function:

$$L_{Ridge} = ||y - Xw||_2^2 + \lambda||w||_2^2, \tag{4.4}$$

where λ is a positive number that penalizes large weights in w (often referred to as the regularization parameter). Taking the derivative of the objective function, Equation (4.4), with respect to w, one can find that the optimal solution of ridge regression is (Saunders et al. 1998):

$$w = (X'X + \lambda I_N)^{-1}X'y,$$

[6]In our setting, problems of multicollinearity can arise since many of the selected words are correlated with each other. Therefore, ordinary least squares is not be applicable.

where I_N is the $N \times N$ identity matrix. By introducing the regularization parameter λ, the ridge regression can reduce the variance of the estimate at the cost of increasing training errors. In other words, the regularization parameter λ balances the trade-off between the bias and variance of the estimate. In practice, one can use cross-validation techniques to find the optimal λ that minimizes the cross-validation errors (Plutowski 1996).[7] In the next section, we discuss kernel ridge regression which improves the computational efficiency of ridge regression by means of 'kernel trick'.

4.3.2 Kernel Ridge Regression

The ridge regression method described above identifies only linear relations between variables. To expand our forecasting model to include non-linear relations, we employ kernel ridge regression (KRR).[8]

Kernel ridge regression maps the set of predictor variables into a high-dimensional (possibly infinite-dimensional) space in such a way that the sought relations can be presented in a linear form (called the *kernel trick*). To avoid over-fitting, a forecasting model is estimated in this high-dimensional space using the penalty term λ. Desired properties of the model are achieved by choosing a kernel in a convenient way to prevent actual calculations in high-dimensional space. KRR overcomes computational difficulties that are encountered in standard ridge regression when the number of predictors is large relatively to the number of observations.[9]

Kernel Ridge Regression (KRR), Basics

[7] To determine the optimal parameter λ, we use a 10-fold cross validation, i.e, we randomly partition a learning sample into ten subsets, estimate λ on one subset, and validate it on the other nine subsets. This process is then repeated then times (i.e, ten folds), with each of the ten subsets used only once for validation.

[8] Kernel models are widely used in machine learning for regression and classification problems because of their flexibility, increased predictive accuracy, and high-dimensional data analysis. See, for example, Poggio and Girosi 1990; Vapnik 1998; Solkopf and Smola 2001 and Shawe-Taylor and Cristianini (2004).

[9] Forecast errors for text-enhanced (i.e., high dimension) models are ten times larger using OLS over KRR, whereas there is virtually no difference in quant-only (low dimension) models (results are not tabulated). This results emphasizes the importance of using regularized methods when the number of predictors is large and predictors are not independent with each other.

A *kernel* is a function k that for all x and z from some input space X satisfies

$$k(x, z) = \langle(\phi(x), \phi(z)\rangle,$$

where ϕ is the mapping from X to a (dot product) feature space F ($\phi : x \to \phi(x) \in F$).

For example, consider a two-dimensional predictor matrix $X \in \mathbb{R}^2$ together with the feature map

$$\phi : x = (x_1, x_2) \to \phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \in F = \mathbb{R}^3.$$

Then the prediction function in F is:

$$w_1x_1^2 + w_2x_2^2 + w_3\sqrt{2}x_1x_2.$$

In this example, ϕ maps the data from a two-dimensional input space to a three-dimensional feature space in a way that linear relations in the feature space correspond to quadratic relations in the input space. The dot product composition of the feature map can be written as follows:

$$\begin{aligned}\langle(\phi(x), \phi(z)\rangle &= \langle(x_1^2, x_2^2, \sqrt{2}x_1x_2), (z_1^2, z_2^2, \sqrt{2}z_1z_2)\rangle \\ &= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 \\ &= (x_1z_1 + x_2z_2)^2 = \langle x, z\rangle^2.\end{aligned}$$

Here the function $k(x, z) = \langle x, z\rangle^2$ is a kernel function. It means that instead of explicitly evaluating the coordinates of the projection of two points into the feature space, one can simply compute the dot product between them. This results in tremendous computational savings and efficiency.

Now we can easy generalize simple ridge regression to kernel ridge regression using the kernel trick. All the data points are replaced with the elements of the corresponding feature space $x_i \to \phi_i = \phi(x_i)$ induced by a kernel k, where $k(x_i, x_j) = \langle(\phi(x), \phi(z)\rangle$. The

prediction of kernel ridge regression given a new observation x can be written as (Saunders et al. 1998):

$$g(x) = (X'\alpha x)' = y'(K + \lambda I)^{-1}k,$$

where

$$K = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_2) & ... & k(x_1, x_L) \\ k(x_2, x_1) & k(x_2, x_2) & ... & k(x_2, x_L) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_L, x_1) & k(x_L, x_2) & ... & k(x_L, x_L) \end{pmatrix}$$

and

$$k = \begin{pmatrix} k(x_1, x) \\ k(x_2, x) \\ \vdots \\ k(x_L, x) \end{pmatrix}.$$

It is important to note that once the kernel function is known and the kernel matrix can be calculated, there is no need to access feature vectors. Taking all these elements together, kernel ridge regression provides a powerful framework for estimating nonlinear relations in a high-dimensional environment.

Our Application

We use the following kernel functions for the **quantitative** data

$$\begin{aligned} k(x_1, x_2) &= x_1'x_2 + c && \text{(Linear)} \\ k(x_1, x_2) &= (c + ax_1'x_2)^b && \text{(Polynomial, degree } b\text{)} \\ k(x_1, x_2) &= e^{-||x_1 - x_2||^2/2\sigma^2} && \text{(Gaussian)} \end{aligned}$$

The linear kernel is the simplest kernel function. It is given by the dot product $\langle x_1, x_2 \rangle$ plus an optional constant c. KRR with the linear kernel is equivalent to the ordinary RR. The polynomial kernel corresponds to a mapping for which $k(x_1, x_2)$ consists of all polynomials in the elements of x of degree b; parameters are the slope a, the constant term c, and the

polynomial degree b. Note that when $b = 1$ and $a = 1$, KRR is equivalent to RR. The parameter σ of the Gaussian kernel impacts model performance significantly; if σ is too high or too low, the exponential function will behave almost linearly or be sensitive to noise in the training data, respectively.

A kernel for textual data is constructed separately.[10] Using the bag-of-words approach, each MD&A document in the dataset can be represented as a row vector:

$$\phi(d) = (tf(t_1, d), tf(t_2, d), ..., tf(t_N, d)) \in \mathbb{R}^N,$$

where $tf(t_i, d)$ is the frequency of the term t_i in the document d.

The advantage of the bag-of-words representation its simplicity (i.e., a numerical count of frequencies), but it has several shortcomings. It neglects the order and semantic content of the words. To address some of these issues, consider the following transformation of the bag-of-words representation:

$$\tilde{\phi}(d) = \phi(d)S,$$

where S is the *semantic matrix*. After this transformation, the corresponding kernel is:

$$\tilde{k}(d_1, d_2) = \phi(d_1)SS'\phi(d_2)' = \tilde{\phi}(d_1)\tilde{\phi}(d_2)'. \tag{4.5}$$

Different choices of semantic matrix S lead to different types of kernels for text analysis. One might consider S to be the matrix that assigns weights to the words. In Section 4.1.2, we discussed six term weighting rules that can modify a simple BOW representation of texts. The goal is to assign higher weights to less common words (overall, based on industry, based on company), assuming that frequently used words are less informative. Therefore, the semantic matrix S modifies the simple bag-of-words representation of texts with the inverse document frequency measure, idf. Given the (1)-(6) measures in Section 4.1.2, the semantic matrix S is diagonal with entries:

[10] It is possible to use standard kernels (i.e., linear, polynomial, and Gaussian) for textual data. For example, a polynomial kernel over the bag-of-words representation is: $k(d_1, d_2) = (\langle \phi(d_1), \phi(d_2) \rangle + 1)^b$.

$S_{ii} = idf(i)$ for weighting rules (1) and (2);

$S^k_{ii} = idf_k(i)$ for weighting rules (3) and (5);

$S^c_{ii} = idf_c(i)$ for weighting rules (4) and (6).

Once the semantic matrix is determined, we can construct a text kernel. For example, a kernel for term weighting (1) simply computes the dot product:

$$\tilde{k}(d_1, d_2) = \phi(d_1)SS'\phi(d_2) = \sum_t idf(i)^2 tf(i, d_1)tf(i, d_2). \tag{4.6}$$

Combination of different kernels

To predict future earnings using both numerical and textual data, we need to combine a selected standard kernel for quantitative data with the constructed kernel for qualitative data. Such algebraic operations as addition, multiplication and exponentiation retain key properties of kernels. Thus, one can build new, more powerful kernels from the existing ones (Lanckriet et al. 2004). For example, given two kernels $k_1(x, z)$ and $k_2(x, z)$ with feature mappings ϕ_1 and ϕ_2, it is possible to create new kernels that are additive, multiplicative, and more. For example,

$$\begin{aligned} k(x, z) &= k_1(x, z) + k_2(x, z), && \text{(additive)} \\ k(x, z) &= \alpha k_1(x, z) + (1 - \alpha)k_2(x, z), \alpha \in [0, 1] && \text{(convex combination)} \end{aligned}$$

To understand the intuition for text enhancements, we use linear ridge regression. Consider Array Biopharma's (Ticker: ARRY) data: ROE in 2003 is -23% and ROE in 2004 is -39%. Using 2Q, the forecast of ROE for 2004 is -13% (and thus the forecast error is large). In the text-enhanced model, the stem *collabor* has a negative sign, and ARRY mentions collaborate(s), collaboration, and collaborator 70 times in their MD&A section. This pushes the forecast in the *ex post* correct direction. In contrast, the stem *favor* has a positive coefficient (of approximately the same magnitude as the coefficient on *collabor*) but ARRY only mentions the word favorable twice. Therefore, it contributes little towards

increasing the forecast (in the wrong direction) relative to the quant-only model. While a positive sign on 'favorable' may not be surprising, we would have no reason to expect a negative sign on 'collaborate' (and its variations). This emphasizes the need to determine the value of a word, based on its usage, not on a pre-specified notion of its tone. In addition to adding to accuracy, we may generate new theories based on the estimated signs of significant coefficients, which might not have otherwise emerged.

Table A.4 lists the words most commonly selected for earnings forecasting over the time period 1999-2010. Based on LMFSD classifications, negative words account for more than half of the fifty most commonly selected words while positive words make up only one-fourth of the top 50 (which is consistent with much larger proportion of negative words in the LMFSD). However, we note that while typically classified as negative, the word 'decrease' may be positive if placed in front of 'costs.' Coefficients of over 28% of the words have signs opposite from those predicted by their category (for example, negative words 'devalue', 'unfounded', and 'disapprove' have positive signs whereas positive words 'achieve', 'integrity' and 'assure' have negative signs). Figure A.3 shows changes in normalized frequencies of word categories and single words over time. Greater variability across normalized frequencies of individual words suggests potential changes in the value and informativeness of each word. The FASB's 2001 rule on asset impairments in 2001 and changes to FASB codification (2007-2009) resulted in an increase use and relevance of the term 'impairment' during our sample period (Panel (b), Figure A.3). In contrast, the normalized frequencies of words 'loss', 'gain' and 'effective' have not varied significantly, despite large economic shocks in our sample period (Panels (b) and (c), Figure A.3).

[Table A.4 about here.]

[Figure A.3 about here.]

Studies that use predetermined text categories implicitly assume that all words in the category are equally informative. However, words that measure positive or negative tone

can vary significantly in their strength and frequency (e.g., catastrophic vs. reluctant, exceptional vs. effective). Adding normalized category counts for the six LMFSD categories increases the adjusted-R^2 of a regression of future ROE on current ROE negligibly; adding in normalized word counts for as few as 10 selected words increases the adjusted-R^2 ten times more than adding in category counts.[11] This further supports using an approach that determines the value of each word's occurrence individually and without predisposition.

4.3.3 Lasso Regression

Although both ridge regression and kernel ridge regression are powerful forecasting models, they do not remove redundant or irrelevant variables. As discussed in Section 4.3.1, ridge regression shrinks the coefficients and keeps all variables in the model. Therefore, to increase the predictive power of the model, one has to consider applying variable selection techniques before the estimation. Adding this additional step to the forecasting process may be costly, therefore, other methods should be considered.

Tibshirani (1996) proposes a new regression regularization technique, called the lasso. The lasso is a least squared method with an L_1-penalty on the regression coefficients. Unlike ridge regression which keeps all the variables in the model, lasso regression shrinks some coefficients and sets other coefficients to 0. Hence, the lasso performs both the variable selection and ridge regression simultaneously. Stated formally, in contrast to ridge regression, lasso regression minimizes the following loss function:

$$L_{Lasso} = \frac{1}{2N}||y - Xw||_2^2 + \lambda||w||_1, \tag{4.7}$$

where N is the number of observations and λ is the nonnegative regularization parameter (i.e., penalty term). As the penalty term λ increases, the number of coefficients set to 0 in w increases. Tibshirani (1996) analyzes the prediction performance of the variable selection, lasso, and ridge regression and finds that none of the methods is always superior to the

[11] Details on these (untabulated) results are available from authors.

other two. Using a series of simulated data, Tibshirani (1996) finds that subset selection is superior when the predictor set includes a 'small number of large effects'. The lasso does best when the predictor set includes 'small to moderate number of moderate-sized effects'. Finally, ridge regression does best when the predictor set includes 'large number of small effects'.

To summarize this section, there is no uniform answer to what model estimation technique works best in high-dimensional settings where problems of multicollinearity are practically inevitable. However, as we discussed in this section, all three techniques, ridge, kernel ridge, and lasso, offer powerful solutions in modern data analysis. Figure A.1 summarizes the estimation timeline of study. Figure A.2 shows the process of incorporating MD&A textual content into forecasting models.

[Figure A.1 about here.]

[Figure A.2 about here.]

Chapter 5

Results

This chapter presents the main findings of the dissertation. In Section 5.1, we report the descriptive statistics for the variables used in our analyses. In Section 5.2, we introduce forecast accuracy measures used to measure the performance of quantitative and text-enhanced models. In Section 5.3, we discuss statistical tests used to test the change in forecast accuracy after adding MD&A texts. In Section 5.4, we report the results of out-of-sample predictions across all company-year observations. In Section 5.5, we identify those firms for which textual information in MD&A is the most informative and report forecast accuracy across time time periods. In Section 5.6, we compare the forecasts of text-enhanced statistical models to analysts' consensus forecasts. Finally, in Section 5.7, we discuss the robustness tests of the study.

5.1 Descriptive Statistics

Table A.3 provides descriptive statistics of both qualitative and quantitative variables for the 23,976 firm-years (5,436 unique firms) in our sample.[1] The mean (median) of *ROE* is 0.060 (0.089). The mean (median) of operating income scaled by average owners' equity,

[1]To estimate our forecasting models, we make do with as few as one observation per firm (but use four if available) to reduce the likelihood of eliminating smaller, less mature firms from our sample.

OPINC, is 0.088 (0.115) and the mean (median) of non-operating income scaled by average owners' equity, *NOPINC*, is -0.029 (-0.028). Non-operating income contains recurring non-operating items like interest, but also non-recurring items like gains and losses, and comprises a significant portion of total income. The mean (median) of firm size (as measured by the natural logarithm of market capitalization), *SIZE*, is 5.889 (5.930) with standard deviation 2.003, suggesting that our sample covers small, medium, and large firms. The magnitudes are similar to those reported in Banker and Chen (2006).

Turning to MD&A characteristics, the mean (median) of total words in the MD&A section is 5,074 (4,447) and approximately 10% of MD&A words are in the LMFSD, indicating financial sentiment.[2] The length of the MD&A section tends to increase over time, and our average measures of FOG and LENGTH are consistent with previous research (see for e.g., Li (2008)).

[Table A.3 about here.]

5.2 Measures of Forecast Accuracy

We seek to evaluate whether a large, text-enhanced model provides superior forecasts to a parsimonious quantitative model that uses several accounting-based variables. In our tests, the text-enhanced model nests quantitative model, that is, it is equivalent to quantitative model when the parameters on all text variables are set to zero. The most commonly used statistic for comparing predictions from nested models is mean squared prediction error (MSPE) (Clark and West, 2007). If text variables extracted from the MD&A section do not help in prediction, the MSPE of quantitative models should be smaller than text-enhanced models because the parsimonious, quantitative model gains efficiency by setting coefficients on text variables equal to zero, while the text-enhanced model may introduce

[2] Loughran and McDonald (2011) report that their dictionaries account for around 5% of words in MD&A. We get a higher percentage (10%) because we perform stemming on both MD&A texts and LMFSD words and delete all stop words in texts.

noise.

Let $\widehat{ROE}_{t+1}^{Q}$ denote the period t forecast of *ROE* in period $t+1$ from the quantitative (Q) model. Let $\widehat{ROE}_{t+1}^{T}$ denote the period t forecast of *ROE* in period $t+1$ from the text-enhanced (T) model. The corresponding period $t+1$ squared prediction errors (SPE) of quantitative and text-enhanced models are $(ROE_{t+1} - \widehat{ROE}_{t+1}^{Q})^2$ and $(ROE_{t+1} - \widehat{ROE}_{t+1}^{T})^2$, respectively. Let N be the number of out-of-sample forecasts. Then, the mean squared prediction errors from Q and T models are defined as:

$$MSPE^{Q} = \frac{1}{N}\sum\left(ROE_{t+1} - \widehat{ROE}_{t+1}^{Q}\right)^2;$$
$$MSPE^{T} = \frac{1}{N}\sum\left(ROE_{t+1} - \widehat{ROE}_{t+1}^{T}\right)^2.$$

If there is no improvement in accuracy from adding textual information to a forecasting model then $MSPE^{Q} = MSPE^{T}$. Alternatively, the text-enhanced model has a smaller *MSPE* than quantitative model. Following Clark and West (2007), we adjust the difference between the MSPEs of the Q and T models for the noise associated with the larger model's forecast. Clark and West (2007)'s adjustment ensures reliable forecast comparisons of nested models. We define adjustment factor $J = \frac{1}{N}\sum\left(\widehat{ROE}_{t+1}^{Q} - \widehat{ROE}_{t+1}^{T}\right)^2$ and calculate the accuracy improvement as

$$AI = MSPE^{Q} - (MSPE^{T} - J),$$

where a positive value of *AI* indicates text-enhanced models are more accurate than quantitative models.

Although our primary evaluation criteria is mean-squared error, we also use mean absolute prediction error (APE) for robustness (results are all quantitatively the same). The absolute prediction error (*APE*) is defined as the absolute value of the difference between

actual and predicted *ROE* values, or

$$APE = \left| ROE_{t+1} - \widehat{ROE}_{t+1} \right| .$$

5.3 Statistical Tests

We use one sided t-test and Wilcoxon sign-rank test to analyze the difference in mean and median squared prediction errors.[3] We use four prior years to estimate the coefficients of the Q and T models, and therefore 1999 becomes our first year of out-of-sample forecasts.

In addition to statistical significance, we calculate the economic significance of the forecasting differences following Fairfield et al. (1996). The economic significance tests compare the percentage of observations in which text increases forecast accuracy by more than 5% to the percentage of observations in which text decreases forecast accuracy by more than 5%. This presumes that investors are indifferent to forecast improvements of less than 5% and we focus on only those observations where the difference across models is significant enough to matter. A binomial test determines whether the proportion of firms with increased forecast accuracy after adding text is significantly larger than the proportion of firms with decreased accuracy.

5.4 Accuracy Improvements from Adding Text

Existing work on textual analysis in accounting and finance has focused on the association between the amount, tone, and complexity of qualitative disclosures and future earnings or returns (see Chapter 2). However, as discussed earlier, in-sample explanatory power is not equivalent to out-of-sample predictive power. Existing earnings forecasting models

[3]The use of one sided t-tests is the most widely accepted approach for comparing the accuracy of nested models (see Clark and McCracken (2013)); whereas Wilcoxon sign-rank test is a non-parametric paired test that does not assume normality as *t*-test does and is less susceptible to outliers.

use only quantitative financial information, ignoring the vast amounts of qualitative disclosures provided by managers and regulated by the SEC. Therefore, it is not clear whether textual content in the MD&A section of annual reports has any predictive power for future earnings. In Section 3.2, we introduced two different ways of incorporating text into a forecasting model: one based on text categories, one based on detailed text analysis. In Section 4, we introduced three different methods of estimating statistical models with many predictors: Ridge Regression, Kernel Ridge Regression, and Lasso Regression. We also discussed benefits of using a feature selection prior to estimation to reduce the number of predictor variables. In this section, we discuss out-of-sample predictive accuracy of traditional quantitative models and those that incorporate text using these different approaches.[4]

First, we start with analyzing which method works better for incorporating text in earnings predictions. Table A.5 provides descriptive statistics of mean squared prediction errors of our quantitative benchmarks (Q) and category-enhanced (C) models over the whole sample period using Ridge Regression (RR, Panel A), Kernel Ridge Regression (KRR, Panel B), and Lasso Regression (LR, Panel C) (MSPEs are aggregated across firms and over time). The reported results are consistent across two models (Model 1 and Model 2), therefore, we focus on discussing Model 2 only. The mean (median) prediction errors for Model 2 are 0.041 (0.007) for both Q-model and C-models using RR. We get mean (median) prediction errors of 0.040 (0.005) using KRR for Q- and C- models. Finally, using LR, we get the mean (median) prediction error of 0.040 (0.005) for Q-model and 0.040 (0.004) for C-model. Enhancing quantitative models with six text categories based on LMFSD classification does not improve forecasting accuracy at all in either Model 1 or Model 2. Therefore, we cannot reject the null hypothesis of no predictive value from text categories. The lack of predictive power of categories may be surprising given that the correlations

[4]We use cross-validation to estimate the penalty term λ for each of estimation methods (see Chapter 4). Our results are qualitatively similar across estimation methods and forecast models, therefore, we report the results of using Kernel Ridge Regression for estimation.

between future ROE and the normalized frequencies of words in the six categories are significant and negative (positive) for the 'negative' and 'litigious' ('positive,' 'modal strong,' and 'modal weak') categories. This result highlights the nature of the difference between out-of-sample forecasting and in-sample explanatory models.

[Table A.5 about here.]

Since Table A.5 demonstrates that simple word categories do not improve earnings forecasts, the next step is to consider more detailed text-enhanced models. Each quantitative earnings forecast model is enhanced with normalized frequencies of individual words, so that a statistical model can estimate the relevance of words included. This approach results in a large number of predictor variables which are correlated. As discussed in Chapter 4, there several ways to deal with high-dimensionality of predictor space. We first discuss the results of estimating quantitative and text-enhanced models using Ridge Regression, Kernel Ridge Regression, and Lasso Regression and no feature selection. Recall that both RR and KRR methods shrink the coefficients estimates, but keep all the predictors in the model. In contrast, LR performs both estimation and feature selection (i.e., some parameters are set to 0 during the estimation). Table A.6 reports the results of using a more detailed way to incorporate text into forecasts with no feature selection prior to the estimation. As before, the reported results are consistent across two models (Model 1 and Model 2) and we focus on discussing Model 2 only. The mean (median) prediction errors for Model 2 are 0.040 (0.006) for Q-model and 0.038 (0.005) for T-model using Ridge Regression. In relative terms, the T-model decreases the average prediction error of Q-model by 5%. The results are statistically significant at 1% using both t-test and sign-rank test. Also, text increases the accuracy of forecasts by more than 5% for 53% of firms and decreases the accuracy of forecasts by more than 5% for 37% of firms. The 16% difference is significant using the binomial test. Results for the remaining two estimation methodologies are similar. Using KRR, we get the mean (median) prediction error of 0.040 (0.005) for Q-model and 0.038 (0.004) for T-model. Finally, using LR, we get the mean (median) prediction error of 0.041 (0.006) for Q-model and 0.040 (0.005). All the results are statistically and economically

significant. Overall, we observe that there are no significant differences in forecast accuracy across the considered estimation methodologies (Ridge, Kernel Ridge and Lasso regressions).[5]

[Table A.6 about here.]

Given we find significant accuracy improvements using individual words, but not categories, the aggregation of many words into one category (some of which may be relevant and others irrelevant for earnings prediction) leads to a measure that is too noisy for prediction. When evaluated on a word-by-word basis, our results indicate that MD&A is indeed relevant in the sense of predictive ability; text-enhanced models lead to more accurate forecasts than quantitative models.[6,7]

Next we use a stepwise regression feature selection (with 5% p-value threshold) prior to estimating the models using Ridge Regression, Kernel Ridge Regression, and Lasso Regression. Given that feature selection methods are designed to remove redundant or irrelevant variables, we expect to see larger improvements in accuracy of text-enhanced models. Table A.8 reports the results. On average we have 140 word-based variables selected for earnings prediction using stepwise regression. Although, there are some improvements in accuracy for KRR and LR, we do not observe any major differences in terms of statistical and economic significance. We observe no changes for RR. Using KRR, the adjusted

[5] In Section 5.2, we discussed a procedure of adjusting the squared errors of larger text-enhanced models for the noise introduced by text variables (see Clark and West (2007) for more details). Table A.7 reports 'raw' squared prediction errors of quantitative and text-enhanced models, i.e., SPE of text-enhanced models are not adjusted for the noise introduced by textual variables. Although unadjusted statistics of text-enhanced models are more conservative (i.e., unadjusted squared errors are generally higher than adjusted squared errors), we still get statistical and economic improvements in forecast accuracy after adding MD&A texts. The t-statistic (z-statistic) decreases from 12.6 to 5.13 (27.3 to 12.1) for Model 2 using Ridge Regression when compared to results reported in Table A.6. The difference in groups of $>5\%$ vs. $<5\%$ improvements using the binomial test is now 7%. We get similar downward changes in statistical tests for Kernel Ridge and Lasso regressions, but all the results are statistically and economically significant.

[6] We replicate our results using return on assets and earnings per share as dependent variables. Text-enhanced models are superior to quant-only models for ROA and EPS forecasts as well.

[7] Our findings are robust across firm size, with both types of models being more accurate for larger firms than smaller firms. While smaller firms are generally harder to predict, the accuracy improvements are consistent across firm size, suggesting that MD&A sections contribute similarly to accuracy improvements for large and small firms.

mean (median) squared prediction error of text-enhanced Model 2 is 0.039 (0.004) which is significantly lower than the squared prediction errors of the quantitative benchmark (p-value$<$0.01). Text-enhanced forecasts are superior 51% of the time and inferior 37% of the time when the accuracy difference across Q and T models is at least 5% (p-value$<$0.01). Using Lasso, we get the adjusted mean (median) squared prediction error of text-enhanced Model 2 of 0.038 (0.004) which is significantly lower than the squared prediction errors of the quantitative benchmark (p-value$<$0.01). Text-enhanced forecasts are superior 53% of the time and inferior 42%. The 11% difference is significant using the binomial test. When we use a smaller threshold for stepwise feature selection (e.g., 1% p-value), our results are practically the same (see Table A.9). We conclude that both Kernel Ridge Regression and Lasso Regression work well in our prediction setting. Therefore, we use Kernel Ridge Regression Results in our further analysis.

[Table A.8 about here.]

[Table A.9 about here.]

It is important to note that reported squared prediction errors of both quantitative and text-enhanced models are likely to be correlated within firms and possibly within years. All statistical tests we use for pair-wise comparisons of quantitative and text-enhanced errors are standard and do not account for such possibility. To draw some inferences about changes in the statistical power of the t-test due to multiple observations per firm and per year, we use regression models with clustered standard errors. More specifically, we regress the difference in squared prediction errors of quantitative and text-enhanced models on a constant and cluster standard errors by firm, by year, and by both firm and year. The resulting t-statistics test the equality of the coefficient to 0, while accounting for possible correlations across firms and years. The (untabulated) results for Model 2 using Ridge Regression are the following: the t-statistics decreases from 12.6 to 11.1 when standard errors are clustered by firm; from 12.6 to 8.4 when standard errors are clustered by year; and from 12.6 to 8 when standard errors are clustered by firm and year. We get similar

changes for other model specifications and estimation methods. These results indicate that the correlations of errors within firms and years reduce but do not eliminate the significant differences in the forecast accuracy across quantitative and text-enhanced models using the t-test. Similar changes are likely to affect the power of the sign-rank test as it assumes independence of observations.

5.5 MD&A Informativeness across Firms and Time Periods

The SEC expects firms to explain uncertainties in their MD&A sections. Intuitively, if firms' financial indicators are strong, adding text into a forecasting model would add little to improve upon quantitative accounting-based measures. On the other hand, if firms' financial indicators are uncertain, adding textual context into the forecasting model may result in superior forecasts. To measure the forecast improvement from adding text, we calculate the pair-wise differences between mean squared prediction errors of a quantitative model and its text-enhanced competitor (we use AI to denote the accuracy improvement). Positive (negative) values of AI indicate that the text-enhanced model generates a lower (higher) level of $MSPE$ and thus is more (less) accurate in forecasting one-year-ahead ROE, as compared to the quantitative benchmark model.

We regress *AI*, the informativeness of MD&A, on firm characteristics; a positive coefficient means that firms with high values of associated variable have more informative MD&A sections. The independent variables of interest are: the magnitude and sign of the change in current and future ROE, accruals, firm size, market-to-book, Z-score, intangibles, audit quality (an indicator equal to 1 if the firm is audited by a Big 4 auditor), the number of business and geographic segments, the FOG index and the length of the MD&A.

We use the S&P quality ranking to control for inherent earnings uncertainty.[8] Table A.10 reports the results of regressing the accuracy improvement on firm characteristics. Standard errors are adjusted for clustering by firm. Column (1) of Table A.10 shows that firms with larger changes in future performance, higher accruals, larger size, lower Z-scores, and lower complexity have more informative MD&A sections. Interestingly, we find that when future changes are negative, disclosures are relatively more informative than when future changes are positive. This may be consistent with firms feeling obliged to provide more informative disclosure to avoid litigation or soften bad news when future changes are expected to be disappointing. We find that firms with positive changes in current performance provide more informative MD&A sections that those that have negative changes. This result is consistent with favorable performance tends being more persistent than negative performance.

To test Hypothesis 3(a), whether text has increased in informativeness following the regulatory reforms, we create an indicator variable *POST* which is equal to 1 if the earnings forecast is in the period 2003-2010 and 0 otherwise. Column (2) of Table A.10 includes the *POST* dummy after controlling for other determinants of accuracy improvements. The coefficient on *POST* is 0.047 (t-statistics of 3.70), indicating the increase in informativeness in the post-regulation period. We create a second indicator variable, *CRISIS* equal to 1 if the forecast is in the period 2007-2009. We include it and interaction terms with the firm's cash position and sector to test Hypothesis 3(b). The coefficient on *CRISIS* is -0.046 (t-statistics of -2.93), indicating that text is less informative in periods of unexpected economic shocks. The coefficients on the interactions terms $CRISIS \times CDSD$ is -0.049 (t-statistics of -2.31) and the coefficient on $CRISIS \times CASH$ is 0.146 (t-statistics of 2.02). These results support our predictions that firms in the consumer discretionary sector and firms with low cash positions are more subject to declines in MD&A informativeness during the crisis period.

[8] We have 19,997 out-of-sample forecasts generated by Q- and T- time-series models (four prior years are used to estimate models in each forecasting window). Adding firm characteristics to our tests of H2 reduces the sample to 16,658 observations.

Figure A.4 plots year-by-year relative percentage differences in mean squared prediction errors of Models 2Q and 2T. Figure A.5 plots year-by-year relative percentage differences in mean absolute prediction errors of Models 2Q and 2T. Positive values indicate text-enhanced model superiority; negative values indicate quantitative model superiority. Text begins to help dramatically in 2003, following new MD&A regulations and continues to improve earnings forecasts through 2006. The benefits of text decrease dramatically in the period 2007 to 2009, when the economy experienced one of the largest unanticipated shocks in history.

[Figure A.4 about here.]

[Figure A.5 about here.]

Two columns of Table A.10 report insignificant negative coefficients on the *FOG* index and document *LENGTH*, suggesting that more readable or shorter MD&A sections do not lead to greater forecasting accuracy. Given that a machine, rather than a human, is determining relevant words in our analyses, more sophisticated language adds no additional processing cost but can possibly differentiate between firms.

[Table A.10 about here.]

Table A.11 shows overall mean and median accuracy levels in each sub-period (*PRE*, *POST*, and *CRISIS*). In the pre-regulation period, text-enhanced models generate a 0.001 (p-value$<$0.01) improvement in accuracy, whereas in the post-regulation period, text-enhanced models improve accuracy by 0.002 (p-value$<$0.01). The economic improvement of forecast accuracy is 2.5% in the pre-regulation period and 12% in the post regulation period. Although both results are statistically significant using the binomial test, the magnitude of improvement from adding text seems to be higher int the post regulation period (12% vs. 2.5% difference). This finding suggests that recent regulatory reforms may have improved the informativeness of MD&A disclosures.

Separating the post-regulation period into the non-crisis and crisis sub-periods, text-enhanced models improve the forecast accuracy by 0.002 in the non-crisis period. In the

crisis period, the accuracy improvement of text-enhanced over quantitative models is 0.001, approximately the same as in the pre-regulation period. These results are consistent with the coefficients on *POST* and *CRISIS* reported in Column (2) of Table A.10.

[Table A.11 about here.]

5.6 Text-enhanced Forecast vs. Analysts' Consensus Forecasts

To test Hypothesis 4(a), we examine the difference in accuracy of text-enhanced and quant-only models for the subset of firms without analyst following (10,924 observations). Results are reported in Panel A of Table A.12. Mean (median) forecast errors from 2Q are 0.041 (0.007) whereas mean forecast errors from 2T are 0.040 (0.006). Both mean and median *AI* are positive and significant (p-value$<$0.01), indicating that text-enhanced models are more accurate. This strongly speaks for the importance of combining narrative disclosures with quantitative information for those firms where no intermediary provides forecasts to improve market expectations. Whether these narrative disclosures are indeed incorporated by the market is an open question for future research.

Turning to the subsample of firms where analyst forecasts are available (13,052 observations), we compare the accuracy of text-enhanced models to the analysts' consensus forecasts.[9]

Panel B of Table A.12 reports the mean (median) absolute prediction error of text-enhanced (T) models and analysts consensus forecasts (A) for high, low, and medium analyst following firms. Consistent with standard economic theory which suggests that analysts are more accurate in more competitive forecasting environments, the forecast errors for high and medium following firms are lower than the forecast errors for small following

[9]Text-enhanced models remain superior to quant-only models for the subset of firms with analyst following, overall and in size or analyst following decile.

firms. The mean difference in SPE between text-enhanced Model 2 and analysts for high [medium] (low) analyst following is 0.002 [-.001] (-0.003). Analysts are superior for firms with high following, but analysts are less or equally accurate than text-enhanced models when following is lower.

By restricting ourselves to annual MD&A data only, we are using only a small fraction of the information available to analysts (e.g., press releases, news articles, industry data, etc.). The amount of information available is generally increasing in firm size.[10] We look at whether the analysts' superiority is related to firm size, to provide some rudimentary evidence on whether analysts are genuinely better at forecasting or whether they simply use more information than our text-enhanced models currently incorporate. Panel C of Table A.12 reports the results of splitting our sample into large, medium and small firms (top, middle and bottom third of *MCAP*). Consistent with prior research, we find that analysts are superior to text-enhanced models for large firms (forecast errors are lower for analysts by 0.003, p-value$<$0.01), but they are not superior for smaller firms (forecast errors are higher by 0.002, p-value$<$0.01).[11] This further justifies efforts to enhance models with qualitative information, as analysts do not necessarily have an advantage in their ability, only in their information set.

[Table A.12 about here.]

[10]The methodology we develop in this dissertation can be used to combine textual information from different firm-related disclosures. For example, incorporating textual from earnings conference calls may provide more timely management perspective on future earnings. Earnings conference call is a less formal source of communication between managers and investors (i.e., managers' prepared remarks and forward-looking statements are probably less screened by lawyers when compared to the SEC reports) and it is largely driven by questions from analysts who are present during the call. One can also use the textual part of analysts' reports to see whether it provides some incremental information about future earnings. Given a substantial evidence in the literature that analysts' forecasts of future earnings are often biased, it may be the case that their textual explanations of derived forecasts could reduce the bias. Finally, firm-related financial news may be used to estimated the public opinion about the financial strength of a firm, its industry position, etc.

[11]Our models are also at a timing disadvantage as the analysts may have a significant portion of first-quarter financial information that we do not utilize in our text-enhanced models.

5.7 Robustness Checks

This section discusses our robustness tests. First, we use two other popular time-series models for forecasting one-year-ahead earnings. The first separates earnings into its cash flows and accrual components (following Sloan (1996)) and the second incorporates cost stickiness with sales declines (following Banker and Chen (2006)). All results are qualitatively similar to those reported in the study. Second, we perform the analysis using return-on-assets (ROA) with average total assets as the scaling variable and earnings per share (EPS) with average shares outstanding (or average shares outstanding multiplied by price) as the deflator. Results are similar to those reported in the study. Third, we consider five different term weighting schemes in calculating word frequencies weights (see Section 4.1.2). Results are similar to those reported in the tables. Fourth, we account for simple negation in texts, i.e., while counting frequencies of LMFSD words, we check whether a word is preceded by negator (no, not, none, neither, never, nobody, don't, haven't, etc.). In this way, we separate different information contents of sentiment words and their negated versions (for example, we record separately frequency counts for "assure" and "cannot assure"). All results remain qualitatively the same. Fifth, we focus our analysis on the content of forward-looking statements in the MD&A section. We extract forward-looking sentences using Li (2010)'s approach. Li (2010) identifies a sentence as forward-looking if it contains any of the following words: *"will", "should", "can", "could", "may", "might", "expect", "anticipate", "believe", "plan", "hope", "intend", "seek", "project", "forecast", "objective",* or *"goal"*. We find that text-enhanced models based on forward-looking statements are superior than quantitative models, although the results are weaker. This suggests that both forward-looking information and historical textual information matter for earnings prediction. Finally, we choose different parameters for the rolling window estimation. When we reduce the number of training years to 2 or 3, all the differences between quantitative and text-enhanced models remain the same, although squared prediction errors of all the models become larger. When we increase the number of training years, the results do

not change. We also estimate all the models using Gaussian and linear kernels for KRR. The resulting absolute prediction errors for models with Gaussian kernel are very similar to those reported in the paper. The linear kernel produces forecast errors of higher magnitude for all the models, but the direction of results does not change.

Chapter 6

Conclusions

In recent years, business data have started being collected at a dramatic pace. With these rich flows of digitized numerical and textual information, accounting and finance professions have started learning new state-of-the-art computational techniques and tools. The major purpose of these techniques is to build better explanatory and predictive models to assist investors, auditors, creditors and other market participants. This dissertation develops unique methodologies for evaluating firm-related texts and financial quantitative information and tests those methodologies and their ability to improve existing performance prediction models. We examine whether textual information in public filings is helpful in assessing future earnings, above and beyond traditional financial variables.

We develop techniques that allow us to combine text with financial variables to come up with explicit firm-level future earnings forecasts. We analyze accuracy improvements in out-of-sample setting and find that text-enhanced forecasting models are significantly and economically more accurate than models using financial variables alone. We find that MD&A sections are more informative for firms with larger changes in future performance, higher accruals, larger size, and lower Z-scores.

The SEC's significant efforts over our sample period to improve the informativeness of MD&A suggest that in earlier periods MD&A may not been as helpful at improving forecasts. When we divide our sample into two sub-periods, 1999-2002 and 2003-2010, we

find that MD&A was significantly less informative in the pre-regulation period. This is one of the first studies to provide some empirical evidence on the success of recent regulatory reforms. During the 2007-2009 financial crisis, MD&A loses much of its informativeness; mostly firms in the consumer staples sector or those that have high cash positions provide more informative informative MD&As during the crisis. This result suggests that although regulatory reforms are designed to improve the quality of disclosures, they cannot eliminate uncertainty caused by large unanticipated economic shocks.

The main motivations for using analysts' forecasts as a proxy for earnings expectations is their superiority to time series models and their availability at relatively low cost. However, many firms are not followed, and for these firms, we find that text-enhanced models are superior to quantitative models. Analysts' superiority differs across firm size and analyst following; analysts lose their superiority over text-enhanced models for firms with less competitive forecasting environments (i.e., low analyst following) and for smaller firms.

Bibliography

Abrahamson, E. and E. Amir (1996). The information content of the president's letter to shareholders. *Journal of Business Finance Accounting 23*(8), 1157–1182.

Antweiler, W. and M. Z. Frank (2004). Is all that talk just noise? the information content of internet stock message boards. *Journal of Finance 59*(3), 1259–1294.

Aue, A. and M. Gamon (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of recent advances in natural language processing (RANLP)*, Volume 1, pp. 2–1. Citeseer.

Bagby, J. W., M. R. Kintzele, and L. K. P (1988). Management discussion and performance: An analytical and empirical evaluation. *American Business Law Journal 26*, 57 – 98.

Bainbridge, S. (2007). *The complete guide to Sarbanes-Oxley: Understanding how Sarbanes-Oxley affects your business*. Adams Media.

Ball, R. and P. Brown (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research 6*, 159–178.

Banker, R. D. and L. T. Chen (2006). Predicting Earnings Using a Model Based on Cost Variability and Cost Stickiness. *The Accounting Review 81*(2), 285 – 307.

Barron, O. E., C. O. Kile, and T. B. O'KEEFE (1999). Md&a quality as measured by the sec and analysts' earnings forecasts*. *Contemporary Accounting Research 16*(1), 75–109.

Bell, R. M., Y. Koren, and C. Volinsky (2008). The bellkor 2008 solution to the netflix prize. *Statistics Research Department at AT&T Research*.

Berry, M. W. and M. Castellanos (2004). *Survey of text mining*. Springer.

Bloomfield, R. J. (2002). The "incomplete revelation hypothesis" and financial reporting. *Accounting Horizons 16*(3), 233–243.

Boiy, E., P. Hens, K. Deschacht, and M.-F. Moens (2007). Automatic sentiment analysis in on-line text. In *ELPUB*, pp. 349–360.

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science 2*(1), 1–8.

Botosan, C. A. (1997). Disclosure level and cost of equity capital. *The Accounting Review 72*(3), 323 – 349.

Bradshaw, M., M. Drake, J. Myers, and L. Myers (2012). A re-examination of analysts superiority over time-series forecasts of annual earnings. *Review of Accounting Studies*, 1–25.

Brockman, P., X. Li, and S. M. Price (2012). Do managers put their money where their mouths are? evidence from insider trading after conference calls. In *Evidence from Insider Trading after Conference Calls (March 15, 2012)*.

Brown, L. D., R. L. Hagerman, P. A. Griffin, and Z. M. E. (1987). Security analyst superiority relative to univariate time-series models in forecasting quarterly earnings. *Journal of Accounting and Economics 9*(1), 61–87.

Brown, S. V. and J. W. Tucker (2011). Large-Sample Evidence on Firms Year-over-Year MD&A Modifications. *Journal of Accounting Research 49*(2), 309 – 346.

Bryan, S. H. (1997). Incremental information content of required disclosures contained in Management Discussion and Analysis. *The Accounting Review 72*(2), 285 – 301.

Buckley, C. (1993). The importance of proper weighting methods. In *Proceedings of the workshop on Human Language Technology*, pp. 349–352. Association for Computational Linguistics.

Bushee, B. J., D. A. Matsumoto, and G. S. Miller (2004). Managerial and investor responses to disclosure regulation: The case of reg fd and conference calls. *The Accounting Review 79*(3), 617–643.

Bushman, R. M. and A. J. Smith (2001). Financial accounting information and corporate governance. *Journal of Accounting and Economics 32*(13), 237 – 333.

Callahan, C. M. and R. Smith (2004). Firm performance and management's discussion and analysis disclosures: An industry approach. *Available at SSRN 588062*.

Chaovalit, P. and L. Zhou (2005). Movie review mining: A comparison between supervised and unsupervised classification approaches. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pp. 112c–112c. IEEE.

Chen, H., P. De, J. Hu, and B.-H. Hwang (2013). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*.

Chen, J., E. Demers, and B. Lev (2013). Oh what a beautiful morning! the time of day effect on the tone and market impact of conference calls. *Working paper*.

Chen, J. V. and F. Li (2013). Estimating the amount of estimation in accruals. *Working Paper*.

Clark, T. and M. McCracken (2013). Chapter 20 - advances in forecast evaluation. In *Handbook of Economic Forecasting*, Volume 2, Part B of *Handbook of Economic Forecasting*, pp. 1107 – 1201.

Clark, T. and K. West (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics 138*, 291–311.

Clarkson, P. C., J. L. Kao, G. D, and Richardson (1999). Evidence that Management Discussion and Analysis (MD&A) is a part of firms overall package. *Contemporary Accounting Research 16*(1), 111 – 134.

Cole, C. J. and C. L. Jones (2004). The usefulness of md&a disclosures in the retail industry. *Journal of Accounting, Auditing & Finance 19*(4), 361–388.

Core, J. E. (2001). A review of the empirical disclosure literature: discussion. *Journal of Accounting and Economics 31*(1), 441–456.

Cutler, D. M., J. M. Poterba, and L. H. Summers (1989). What moves Stock Prices? *Journal of Portfolio Management 15*(3), 4 – 12.

Davis, A. K., J. M. Piger, and L. M. Sedor (2006). Beyond the numbers: An analysis of optimistic and pessimistic language in earnings press releases. *Federal Reserve Bank of St. Louis Working Paper Series* (2006-005).

Dechow, P. M. and W. Ge (2006). The persistence of earnings and cash flows and the role of special items: Implications for the accrual anomaly. *Review of Accounting Studies 11*(2-3), 253–296.

Demers, E. and C. Vega (2008). *Soft information in earnings announcements: News or noise?* Federal Reserve Board.

Draper, N. and H. Smith (1966). *Applied regression analysis*. John Wiley and Sons, New York.

Economist, T. (2010). Data, data everywhere - a special report managing information. *The Economist Newspaper LTD*.

Engelberg, J. (2008). Costly information processing: Evidence from earnings announcements. In *AFA 2009 San Francisco Meetings Paper*.

Fairfield, P. M., R. J. Sweeney, and T. L. Yohn (1996). Accounting classification and the predictive content of earnings. *The Accounting Review 71*(3), 337 – 355.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics 33*(1), 3–56.

Fan, W., L. Wallace, S. Rich, and Z. Zhang (2006). Tapping the power of text mining. *Communications of the ACM 49*(9), 76–82.

Feldman, L., S. Govindaraj, J. Livnat, and B. Segal (2009). Managements tone change, post earnings announcement drift and accruals. *Review of Accounting Studies 15*(4), 915 – 953.

Francis, J. R. and M. D. Yu (2009). Big 4 office size and audit quality. *The Accounting Review 84*(5), 1521–1552.

Frankel, R. and C. Lee (1998). Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics 25*, 283 – 319.

Frazier, K. B., R. W. Ingram, and B. M. Tennyson (1984). A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research 22*(1), 318 – 331.

Fried, D. and D. Givoly (1982). Financial analysts forecasts of earnings: A better surrogate for market expectations. *Journal of Accounting and Economics 4*(2), 85–107.

Gangolly, J. and Y. Wu (2000). On the automatic classification of accounting concepts: Preliminary results of the statistical analysis of term-document frequencies. *The New Review of Applied Expert Systems and Emerging Technologies 6*, 81 – 88.

Gerakos, J. and R. B. Gramacy (2013). Regression-based earnings forecasts. *Chicago Booth Research Paper* (12-26).

Goel, S., J. Gaangolly, S. R. Faerman, and O. Uzuner (2010). Can linguistics predictors detect fraudulent finacial filings? *Journal of Emerging Technologies in Accounting 7*.

Guyon, I. and A. Elisseeff (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research 3*, 1157–1182.

Henry, E. (2006). Market Reaction to Verbal Components of Earnings Press Releases: Event Study Using a Predictive Algorithm. *Journal of Emerging Technologies in Accounting 3*.

Henry, E. and A. J. Leone (2010). Measuring qualitative information in capital markets research. *Papel de trabajo. University of Miami*.

Hoberg, G. and C. Lewis (2013). Do fraudulent firms engage in disclosure herding? *Available at SSRN 2298302*.

Hoerl, A. and R. Kennard (1988). Ridge regression, in encyclopedia of statistical sciences, vol. 8.

Hooks, K. L. and J. E. Moon (1993). A classification scheme to examine Management Discussion and Analysis compliance. *Accounting Horizons 7*(2), 41 – 59.

Jegadeesh, N. and D. Wu (2013). Word power: A new approach for content analysis. *Journal of Financial Economics 110*(3), 712–729.

Jurafsky, D. and H. James (2000). Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

Khurana, I. K. and K. Raman (2004). Litigation risk and the financial reporting credibility of big 4 versus non-big 4 audits: Evidence from anglo-american countries. *The Accounting Review 79*(2), 473–495.

Kothari, S. (2001). Capital markets research in accounting. *Journal of accounting and economics 31*(1), 105–231.

Kothari, S., X. Li, and J. E. Short (2009). The effect of disclosures by management, analysts, and business press on cost of capital, return volatility, and analyst forecasts: a study using content analysis. *The Accounting Review 84*(5), 1639–1670.

Lanckriet, G. R. G., N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. I. Jordan (2004). Learning the kernel matrix with semidenite programming. *Journal of Machine Learning Research 5*, 27 – 72.

Larcker, D. F. and A. A. Zakolyukina (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research 50*(2), 495–540.

Lee, Y.-J. (2012). The effect of quarterly report readability on information efficiency of stock prices*. *Contemporary Accounting Research 29*(4), 1137–1170.

Lehavy, R., F. Li, and K. Merkley (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review 86*(3), 1087–1115.

Lesikar, R. L. and M. P. Lyons (1986). Report Writing for Business. *Homewood, IL: Irwin. 21*.

Lev, B. (2001). *Intangibles: Management, measurement and reporting*. Brookings Institution Press.

Lev, B. and S. R. Thiagarajan (1993). Fundamental Information Analysis. *Journal of Accounting Research 31*(2), 190 – 215.

Li, F. (2006). Do stock markets investors understand the risk sentiment of corporate annual reports? *Working Paper (April), University of Michigan.*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics 45*(2 - 3), 221 – 247.

Li, F. (2010). The information content of forward-looking statements in corporate filings-a nave bayesian machine learning approach. *Journal of Accounting Research 48*(5), 1049–1102.

Li, F., R. Lundholm, and M. Minnis (2012). A measure of competition based on 10-k filings. *Journal of Accounting Research*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies 5*(1), 1–167.

Loughran, T. and B. McDonald (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance 66*(1), 35 – 65.

Loughran, T. and B. McDonald (2013). Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics 109*(2), 307–326.

Loughran, T. and B. McDonald (2014). Measuring readability in financial disclosures. *Journal of Finance*.

Manning, C. D. and H. Schütze (1999). *Foundations of statistical natural language processing*. MIT press.

Matsumoto, D., M. Pronk, and E. Roelofsen (2011). What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. *The Accounting Review 86*(4), 1383–1414.

Merkley, K. (2013). Narrative disclosure and earnings performance: Evidence from r&d disclosures. *The Accounting Review 89*(2).

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review 85*(6), 2107–2143.

Muslu, V., S. Radhakrishnan, K. Subramanyam, and D. Lim (2014). Forward-looking md&a disclosures and the information environment. *Management Science*.

Nelson, K. K. and A. Pritchard (2007). Litigation risk and voluntary disclosure: The use of meaningful cautionary language. *SSRN eLibrary*.

O'brien, P. C. (1988). Analysts' forecasts as earnings expectations. *Journal of Accounting and Economics 10*(1), 53–83.

Ou, J. and S. Penman (1989). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*, 295 – 330.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pp. 79–86. Association for Computational Linguistics.

Pava, M. L. and M. J. Epstein (1993). How good is MD&A as an investment tool? *Journal of Accountancy 175*(3), 51 – 53.

Plutowski, M. E. P. (1996). Survey: Cross-validation in theory and practice. Research report. *Department of Computational Science Research, David Sarnoff Research Center.*.

Poggio, T. and F. Girosi (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science 247*, 978 – 982.

Price, S. M., J. S. Doran, D. R. Peterson, and B. A. Bliss (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance 36*(4), 992–1011.

Richardson, S. A., R. G. Sloan, M. T. Soliman, and I. Tuna (2005). Accrual reliability, earnings persistence and stock prices. *Journal of Accounting and Economics 39*(3), 437–485.

Rogers, J. L., A. V. Buskirk, and S. L. C. Zechman (2011). Disclosure tone and shareholder litigation. *The Accounting Review 86*(6), 2155–2183.

Roll, R. W. (1988). R-Squared. *The Journal of Finance 43*(3), 541 – 566.

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management 24*(5), 513 – 523.

Saunders, C., A. Gammerman, and V. Vovk (1998). Ridge Regression Learning Algorithm in Dual Variables. *Proceeds Of the 15th International Conference on Machine Learning*, 515 – 521.

Schroeder, N. and C. Gibson (1990). Readability of managements discussion and analysis. *Accounting Horizons 4*(4), 78–87.

SEC (1998). A plain english handbook: How to create clear sec disclosure documents. *U.S. Securities and Exchange Commission 2*.

Shawe-Taylor, J. and N. Cristianini (2004). Kernel Methods for Pattern Analysis. *Cambridge: Cambridge Univ. Press.*.

Shiller, R. J. (1981). Do stock prices move too much to be justified by subsequent changes in dividends? *American Economic Review 71*(3), 421 – 436.

Shmueli, G. (2010). To explain or to predict? *Statistical Science 25*(3), 289–310.

Sloan, R. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting Review 71*, 289 – 315.

Solkopf, B. and A. J. Smola (2001). Learning with Kernels: Support vector machines, regularization, optimization, and beyond. *Cambridge: The MIT Press.*.

Sun, Y. (2010). Do MD&A Disclosures Help Users Interpret Disproportionate Inventory Increases? *The Accounting Review 85*(4).

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics 37*(2), 267–307.

Tetlock, P. C. (2007). Giving Content to investor sentiment: The role of media in the stock Market. *The Journal of Finance 62*(3), 1139 – 1167.

Tetlock, P. C., M. Saar-Tsechansky, and S. Macskassy (2008). More than words: Quantifying language to measure firms fundamentals. *The Journal of Finance 63*(3), 1437 – 1467.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Vapnik, V. (1998). Statistical Learning Theory. *New York: Wiley*.

You, H. and X. Zhang (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting Studies 14*(4), 559 – 586.

Zho, J., D. P. Foster, R. A. Stine, and L. H. Unga (2006). Streamwise Feature Selection. *Journal of Machine Learning Research 7*, 1861 – 1885.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Appendix A

Tables and Figures

Table A.1: Textual Analysis of Corporate Disclosures - Literature Review.

Study	*Sample Size*	*Disclosure Type*	*Text Mining Method*	*Short Summary of Results*
Hooks and Moon (1993)		SEC Disclosures	Frequency and Amount of Disclosure	Companies responded to the release of the SEC 1989 guidelines by increasing the level of their disclosures. The disclosure response is even stronger one year after the 1989 SEC reform.
(Frazier et al., 1984)	74	10-K report	Identification of positive or negative narrative themes	Management analysis data in the annual report is useful for predicting the future performance of a firm. No differences in narrative disclosures across the ownership structure of companies.
Pava and Epstein (1993)	25	MD&A	Identification of forward-looking vs. historical statements	Most companies provide good explanations of historical events, very few provide useful forecasts for the future. Companies are biased in favor of correctly identifying positive news, whereas negative news are either omitted or unclearly reported.
Schroeder and Gibson (1990)	40	MD&A, President's Letter	Document readability	MD&A texts are, in general, difficult to read and comprehend.
Bryan (1997)	250	MD&A	Disclosure Index	MD&A disclosures are associated with short-term future performance. MD&A disclosures are positively and associated with analysts' forecasts around the release date of MD&As.
Cole and Jones (2004)	150	MD&A	Disclosure Index	Disclosure-based variables can predict future revenues, earnings, and are associated with contemporaneous stock returns.
Barron et al. (1999)	550	MD&A	MD&A Quality	There is significant association between MD&A quality ratings and analysts' earnings forecasts.
Clarkson et al. (1999)	55	MD&A	Survey on MD&A	MD&A disclosures are useful to sell-side analysts.
Li (2008)	55,719	10-K report	Document Readability	Companies with low earnings tend to disclose information in a more "difficult-to-read" manner, whereas companies with high earnings prepare more readable reports.
Lee (2012)	60,161	10-Q report	Document Readability	Stock prices of firms with longer or less readable 10-Qs react less strongly to the earnings-related information during the short window following the 10-Q release.

Table A.1 : (continued from the previous page)

Study	*Sample Size*	*Disclosure Type*	*Text Mining Method*	*Short Summary of Results*
You and Zhang (2009)	24,269	10-K report	Document Readability	There are unusual stock price and trading volume movements around the 10-K filing dates. The information complexity of 10-K filings causes stronger under-reaction of investors.
Miller (2010)	12,771	10-K report	Document Readability	More complex reports are significantly associated with lower levels of aggregate trading volume (driven by a reduction in small investor trading volume).
Lehavy et al. (2011)	33,704	10-K report	Document Readability	Analyst following, the amount of effort incurred to generate their reports, and the informativeness of analysts' reports are greater for firms with less readable 10-Ks.
Loughran and McDonald (2014)	66,707	10-K report	Document Readability	Larger 10-K file sizes have significantly higher post-filing date abnormal return volatility, higher absolute standardized unexpected earnings (SUE), and higher analyst dispersion.
Li (2006)	34,180	10-K report	Risk Sentiment	An increase in the number of risky words in annual reports is strongly associated with lower future earnings and stock returns.
Loughran and McDonald (2011)	37,287	10-K report	Tone/Sentiment	The study develops financial sentiment dictionaries. Some of the word lists are related to market reactions around the 10-K filing date, unexpected earnings, subsequent stock return volatility, and events such as accounting fraud or reported material weaknesses in accounting controls.
Feldman et al. (2009)	153,988	10-K (Q) report	Tone	The tone change is associated with both short-window return around the filing date and the drift returns in the post-filing period.
Jegadeesh and Wu (2013)	45,860	10-K report	Tone	The study develops a new measure of document tone based on market reactions. The proposed tone measure is strongly related to filing date returns for both positive and negative word lists, unlike findings in prior studies that only negative words matter.

Table A.1 : (continued from the previous page)

Study	*Sample Size*	*Disclosure Type*	*Text Mining Method*	*Short Summary of Results*
Abrahamson and Amir (1996)	1,300	President's Letters	Tone	The relative negative content of a letter is strongly negatively associated with past and future performance, strongly negatively associated with past and current annual returns, and weakly negatively associated with future returns.
Li (2010)	145,479	MD&A, 10-K & 10-Q	Tone	When managers are more optimistic in their forward-looking statements, future performance is indeed better.
Merkley (2013)	22,482	R&D disclosure, 10-K	Tone, Readability, and Detail	Current earnings performance is negatively associated with the quantity of narrative R&D disclosures. managers use R&D disclosures to provide relevant information rather than obfuscate.
Sun (2010)	568	MD&A, 10-K	Tone, Readability, and Detail	MD&A disclosures on disproportionate inventory increases, if provided, extend financial statement information and help predict future firm performance.
Muslu et al. (2014)	44,708	MD&A, 10-K	Amount of FLS	Firms make more forward-looking MD&A disclosures to improve the information efficiency of stock prices with respect to accounting earnings. The results are stronger for operations-related forward-looking disclosures, disclosures made by loss firms, and disclosures that are made prior to 2000.
Botosan (1997)	122	10-K report	Amount of FLS	Greater disclosure of firms with a low analyst following is associated with a lower cost of equity capital.
Kothari et al. (2009)	$>$100,000	SEC, Analysts, Media	Tone	If disclosures have positive sentiment, the firm's risk declines. In contrast, disclosures with negative sentiment lead to significant increases in risk measures.
Goel et al. (2010)	$<$1,000	10-K report	Disclosure Style and Content	Adding linguistic features to the analysis improves the overall effectiveness of fraud detection.

Table A.1 : (continued from the previous page)

Study	*Sample Size*	*Disclosure Type*	*Text Mining Method*	*Short Summary of Results*
Hoberg and Lewis (2013)	49,039	MD&A, 10-K	Similarity of Disclosure, Tone	Firms use complexity to potentially conceal fraudulent actions, and these firms often use uncertain, litigious, and speculative words.
Nelson and Pritchard (2007)	53,315	10-K report	Cautionary Language	Firms facing greater litigation risk use more cautionary statements in their disclosures.
Brown and Tucker (2011)	28,142	MD&A, 10-K	MD&A Updates	Firms with larger economic changes modify their MD&A more often than those with smaller economic changes. MD&A modification scores have declined in recent years, while MD&A disclosures have become longer.
Loughran and McDonald (2013)	1,887	Form S-1	Tone	Initial Public Offerings with high levels of uncertain text have higher first day returns, absolute offer price revisions, and subsequent volatility.
Demers and Vega (2008)	21,580	Earnings Announcement	Tone	Unexpected net optimism in managers' language affects abnormal returns around announcement periods and predicts post earnings announcement drift.
Davis et al. (2006)	23,400	Earnings Announcement	Tone	There is a significant positive (negative) association between levels of optimistic (pessimistic) tone in earnings press releases and future return on assets.
Henry (2006)	441	Earnings Announcement	Tone, Length, Numerical Intensity, and Linguistic Complexity	Inclusion of a broad range of numerical financial variables does not enhance predictive accuracy of market returns, whereas inclusion of verbal variables does.
Price et al. (2012)	2,880	Earnings Conference Call	Tone	The linguistic tone of earnings conference call is a significant predictor of abnormal returns and trading volume.
Engelberg (2008)	51,207	Earnings Announcement	Complexity, Readability	The textual information contributes uniquely to the PEAD phenomenon. The more confusing the textual information is, the more slowly it is reflected in stock prices.

Table A.1 : (continued from the previous page)

Study	*Sample Size*	*Disclosure Type*	*Text Mining Method*	*Short Summary of Results*
Rogers et al. (2011)	165	Earnings Announcement	Tone	Sued firms use substantially more optimistic language in their earnings announcements than non-sued firms.
Larcker and Zakolyukina (2012)	29,663	Earnings Conference Call	Deceptive Language	Subsequent restatements are more likely when managers use more references to general knowledge, fewer non-extreme positive emotion words, and fewer third-person plural pronouns.
Brockman et al. (2012)	2,880	Earnings Conference Call	Tone	The positive tone of conference calls predicts insider selling, whereas the negative tone of conference calls predicts insider buying.
Matsumoto et al. (2011)	>10,000	Earnings Conference Call	Amount	Managers tend to provide more disclosure during the introductory session when firm performance is poor, but relatively more information is released during the Q&A session.
Chen et al. (2013)	26,585	Earnings Conference Call	Tone	The tone of the conversations between analysts and managers becomes significantly more negative as the day wears off. More negatively toned conversations are associated with more negative abnormal stock returns during the call and immediately after the call.
Henry and Leone (2010)	29,712	Earnings Announcement	Tone	The context-specific word lists are more powerful than the general word lists in measuring the tone.
Tetlock (2007)	3,709	Financial News	Tone	High levels of media pessimism tend to predict downward pressure on market prices, followed by a reversion to fundamentals.
Tetlock et al. (2008)	< 18,000	Financial News	Tone	Increases in the number of negative words used in WSJ and DJNS columns about S&P 500 firms relative to prior stories predict larger negative shocks to future earnings.
Antweiler and Frank (2004)	>1.5 million	Stock Message Boards	Tone	Linguistic features of messages posted on Yahoo! Finance help predict market volatility.
Chen et al. (2013)	97,070	Social Media	Tone	The views expressed in both articles and commentaries predict future stock returns and earnings surprises.

Table A.2: Sample Creation

Source	Sample Size
EDGAR	
Download all 10-K/10-K405 (1995-2010)	112,084
Extract MD&A section from each 10-K report	102,885
COMPUSTAT	
Download data for all companies (1995-2011)[a]	114,395
Delete observations with missing values in selected variables, negative values of owners' equity, and $\|ROE\| > 1$, $\|Sales\text{-}COGS\|/Sales > 1$, and $\|REV_{t-1} - REV_{t-2}\|/REV_{t-2} > 1$	
Exclude financial firms (SIC 6000s)	45,740
EDGAR & COMPUSTAT	
Match MD&A and COMPUSTAT data	26,322
Delete MD&As with fewer than 250 words[b]	23,976

[a] In 2011, we require only earnings, shares outstanding and book value of equity for the companies in our sample.
[b] We require at least 250 words to appear in the MD&A section to eliminate those 10-K reports in which the MD&A section is only incorporated by reference (i.e., pointing shareholders to another file or the annual report).

Table A.3: Descriptive statistics: 1995-2010.

Variable	*N*	*Mean*	*Median*	*Std. Dev.*	*Q1*	*Q3*
Quantitative, Forecasting						
ROE	23,976	0.060	0.089	0.216	-0.011	0.159
OPINC	23,976	0.088	0.115	0.273	-0.014	0.219
NOPINC	23,976	-0.029	-0.028	0.123	-0.071	0.009
Quantitative, General						
SIZE	23,976	5.889	5.930	2.003	4.481	7.226
ACCR	19,997	-0.158	-0.117	0.307	-0.224	-0.036
INTAN	19,226	0.163	0.082	0.193	0.006	0.264
MTB	19,997	1.307	0.896	1.547	0.499	1.578
Z-SCORE	17,828	4.629	3.188	6.817	1.838	5.343
NBSEG	19,997	0.917	0.693	0.493	0.004	1.386
NGSEG	19,997	0.893	0.693	0.559	0.004	1.386
CDSD	19,997	0.525	-	0.499	-	-
S&P_QR	19,997	0.061	-	0.240	-	-
Text						
FOG	23,976	18.29	17.78	2.536	16.79	19.56
MD&A WORDS	23,976	5,074	4,447	3,494	2,636	6,625
LMFSD WORDS	23,976	505	432	373	242	668
MD&A LENGTH	23,976	8.296	8.399	0.739	7.877	8.798
PERC LMFSD	23,976	0.097	0.096	0.018	0.085	0.108

Variable definitions: *ROE* is net income before extraordinary items divided by the average book value of owners' equity; *OPINC* is operating income after depreciation, net of interest expence, special items, and minority interest, divided by the average book value of owners' equity; *NOPINC* is non-operating income, net of income taxes, divided by the average book value of owners' equity; *SIZE* is the natural logarithm of the market value of firm's equity; *ACCR* is earnings minus cash flow from operations divided by the average book value of owners' equity; *MTB* is the market value of equity scaled by the book value of total assets; *Z-SCORE* is the Altman z-score; *NBSEG* is the logarithm of 1 plus the number of business segments; *NGSEG* is the logarithm of 1 plus the number of geographic segments; *CDSD* is a dummy variable equal to 1 if a company belongs to the consumer discretionary sector and 0 otherwise; *S&P_QR* is a dummy variable equal to 1 if S&P Quality Ranking of a company is B+ and higher and 0 otherwise; *FOG* is the Fog index of the MD&A section; *MD&A WORDS* (*LMFSD WORDS*) is the number of all words in the MD&A section (number of words in the MD&A section that are LMFSD words); and *PERC LMFSD* is the number of LMFSD words in MD&A divided by the total words in MD&A.

Table A.4: List of top 50 words selected in forecasting models

Rank	*Word*	*LMFSD Category*	*Rank*	*Word*	*LMFSD Category*
1	loss	Negative	26	unable	Negative
2	strong	Positive	27	assure	Positive
3	amend	Litigious	28	gain	Positive
4	favorable	Positive	29	unfounded	Negative
5	effective	Positive	30	law	Litigious
6	integrity	Positive	31	resign	Negative
7	collaborate	Positive	32	uncertainty	Uncertainty
8	informative	Positive	33	uncollectable	Negative
9	restructure	Negative	34	achieve	Positive
10	compensatory	Litigious	35	outstanding	Positive
11	will	Modal Strong	36	adverse	Negative
12	alert	Negative	37	cancel	Negative
13	benefit	Positive	38	ascendant	Litigious
14	confess	Negative	39	assumption	Uncertainty
15	deficit	Negative	40	burdensome	Negative
16	decline	Negative	41	court	Litigious
17	delist	Negative	42	default	Negative
18	devalue	Negative	43	defer	Negative
19	disapprove	Negative	44	definitely	Modal Strong
20	impair	Negative	45	revise	Uncertainty
21	overrun	Negative	46	delay	Negative
22	severe	Negative	47	exposure	Uncertainty
23	claim	Litigious	48	imperil	Negative
24	contingency	Uncertainty	49	improve	Positive
25	violate	Negative	50	possible	Modal Weak

Table A.5: Squared prediction errors for Ridge Regression, Kernel Ridge Regression, and Lasso Regression with Text Categories. Forecast Years: 1999-2010, all firms.

Panel A: Estimation Methodology - Ridge Regression.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.007	0.040	0.005
SPE_C-*adj*	0.041	0.006	0.040	0.005
AI	0.000	0.001^{+}	0.000	0.000
Statistics	(1.36)	(1.78)	(1.60)	(1.02)
Econ Improve$^{\diamond}$	40.1/39.0		40.2/38.9	

Panel B: Estimation Methodology - Kernel Ridge Regression.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_C-*adj*	0.041	0.006	0.040	0.005
AI	0.000	-0.000	0.000	-0.000
Statistics	(0.16)	(-0.17)	(0.41)	(-1.41)
Econ Improve$^{\diamond}$	38.3/37.9		36.5/35.9	

Panel C: Estimation Methodology - Lasso Regression.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_C-*adj*	0.041	0.005	0.040	0.004
AI	0.000	0.001	0.000	0.001
Statistics	(1.45)	(1.56)	(1.59)	(1.40)
Econ Improve$^{\diamond}$	36.5/35.9		37.9/37.6	

Q is used to denote quantitative forecasting models, *C* is used to denote forecasting models that incorporate both quantitative variables and text categories. SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.

SPE_C-*adj* is the adjusted squared prediction error from the C-model, calculated as the squared difference between actual and predicted *ROE* and adjusted for the squared difference in ROE predictions from the Q- and C-models.

AI is the accuracy improvement gained from using a larger C-model, calculated as the difference in SPE_Q and SPE_C-*adj*.

$^{\diamond}$ Percentage of observations for which text improves forecast accuracy by 5% or more (first number) and the percentage of observations for which text reduces forecast accuracy by 5% or more (second number).

† indicates that the proportion of observations with improved accuracy exceeds the proportion of observations with reduced accuracy using the binomial test at the 1% significance level.

***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test). Number of observations: 23,976.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$,
Model 1C : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(CategMatr_t) + e_{t+1}$,
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$,
Model 2C : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(CategMatr_t) + e_{t+1}$.

Table A.6: Squared prediction errors for Ridge Regression, Kernel Ridge Regression, and Lasso Regression with No Feature Selection. Forecast Years: 1999-2010, all firms.

Panel A: Estimation Methodology - Ridge Regression and No Feature Selection.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.007	0.040	0.006
SPE_T-*adj*	0.039	0.006	0.038	0.005
AI(Q-T)	0.002^{***}	0.001^{+++}	0.002^{***}	0.001^{+++}
Statistics	(16.3)	(27.6)	(12.6)	(27.3)
Econ Improve$^{\diamond}$	$52.8^{\dagger}$/37.1		$53.1^{\dagger}$/37.8	

Panel B: Estimation Methodology - Kernel Ridge Regression and No Feature Selection.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_T-*adj*	0.040	0.005	0.038	0.004
AI	0.001^{***}	0.002^{+++}	0.002^{***}	0.001^{+++}
Statistics	(15.0)	(21.7)	(12.2)	(20.2)
Econ Improve$^{\diamond}$	$38.9^{\dagger}$/27.3		$38.8^{\dagger}$/28.1	

Panel C: Estimation Methodology - Lasso Regression and No Feature Selection.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.041	0.006
SPE_T-*adj*	0.040	0.005	0.040	0.005
AI	0.001^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(12.1)	(20.6)	(6.83)	(20.2)
Econ Improve$^{\diamond}$	$45.3^{\dagger}$/32.7		$42.8^{\dagger}$/31.1	

SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.
SPE_T-*adj* is the adjusted squared prediction error from the T-model, calculated as the squared difference between actual and predicted *ROE* and adjusted for the squared difference in ROE predictions from the Q- and T-models.
AI is the accuracy improvement gained from using a larger text-enhanced model, calculated as the difference in SPE_Q and SPE_T-*adj*.
$^{\diamond}$ Percentage of observations for which text improves forecast accuracy by 5% or more (first number) and the percentage of observations for which text reduces forecast accuracy by 5% or more (second number).
† Proportion of observations with improved accuracy exceeds the proportion of observations with reduced accuracy using the binomial test at the 1% significance level.
***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test). Number of observations: 23,976.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$,
Model 1T : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(TextMatr_t) + e_{t+1}$,
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$,
Model 2T : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(TextMatr_t) + e_{t+1}$.

Table A.7: Squared prediction errors for Ridge Regression, Kernel Ridge Regression, and Lasso Regression with No Feature Selection. Forecast Years: 1999-2010, all firms.

Panel A: Estimation Methodology - Ridge Regression and No Feature Selection.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.007	0.040	0.006
SPE_T	0.040	0.006	0.039	0.005
AI(Q-T)	0.001^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(7.24)	(14.7)	(5.13)	(12.1)
Econ Improve$^{\diamond}$	$49.0^{\dagger}$/40.8		$48.8^{\dagger}$/42.1	

Panel B: Estimation Methodology - Kernel Ridge Regression and No Feature Selection.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_T	0.040	0.005	0.038	0.004
AI	0.001^{***}	0.002^{+++}	0.002^{***}	0.001^{+++}
Statistics	(12.9)	(18.8)	(10.9)	(17.2)
Econ Improve$^{\diamond}$	$38.1^{\dagger}$/28.2		$37.1^{\dagger}$/28.9	

Panel C: Estimation Methodology - Lasso Regression and No Feature Selection.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.041	0.006
SPE_T	0.040	0.005	0.040	0.005
AI	0.001^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(6.08)	(12.8)	(3.68)	(10.9)
Econ Improve$^{\diamond}$	$43.2^{\dagger}$/34.8		$40.4^{\dagger}$/33.5	

SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.
SPE_T is the squared prediction error from the T-model, calculated as the squared difference between actual and predicted *ROE*.
AI is the accuracy improvement gained from using a larger text-enhanced model, calculated as the difference in SPE_Q and SPE_T-*adj*.
$^{\diamond}$ Percentage of observations for which text improves forecast accuracy by 5% or more (first number) and the percentage of observations for which text reduces forecast accuracy by 5% or more (second number).
† Proportion of observations with improved accuracy exceeds the proportion of observations with reduced accuracy using the binomial test at the 1% significance level.
***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test). Number of observations: 23,976.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$,
Model 1T : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(TextMatr_t) + e_{t+1}$,
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$,
Model 2T : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(TextMatr_t) + e_{t+1}$.

Table A.8: Squared prediction errors for Ridge Regression, Kernel Ridge Regression, and Lasso Regression with Feature Selection at 5%. Forecast Years: 1999-2010, all firms.

Panel A: Estimation Methodology - Ridge Regression and Feature Selection at 5%.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.007	0.040	0.006
SPE_T-*adj*	0.039	0.006	0.038	0.005
AI	0.002^{***}	0.001^{+++}	0.002^{***}	0.001^{+++}
Statistics	(17.9)	(29.4)	(13.9)	(29.4)
Econ Improve$^{\diamond}$	$55.3^{\dagger}$/39.3		$55.5^{\dagger}$/39.6	

Panel B: Estimation Method - Kernel Ridge Regression and Feature Selection at 5%.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_T-*adj*	0.039	0.005	0.039	0.004
AI	0.002^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(17.7)	(23.9)	(12.7)	(22.5)
Econ Improve$^{\diamond}$	$51.1^{\dagger}$/36.7		$50.7^{\dagger}$/36.7	

Panel C: Estimation Method - Lasso Regression and Feature Selection at 5%.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_T-*adj*	0.038	0.004	0.038	0.004
AI	0.003^{***}	0.002^{+++}	0.002^{***}	0.001^{+++}
Statistics	(6.06)	(18.3)	(5.34)	(17.4)
Econ Improve$^{\diamond}$	$53.1^{\dagger}$/41.4		$52.8^{\dagger}$/41.7	

SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.
SPE_T-*adj* is the adjusted squared prediction error from the T-model, calculated as the squared difference between actual and predicted *ROE* and adjusted for the squared difference in ROE predictions from the Q- and T-models.
AI is the accuracy improvement gained from using a larger text-enhanced model, calculated as the difference in SPE_Q and SPE_T-*adj*.
$^{\diamond}$ Percentage of observations for which text improves forecast accuracy by 5% or more (first number) and the percentage of observations for which text reduces forecast accuracy by 5% or more (second number).
† Proportion of observations with improved accuracy exceeds the proportion of observations with reduced accuracy using the binomial test at the 1% significance level.
***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test). Number of observations: 23,976.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$,
Model 1T : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(TextMatr_t) + e_{t+1}$,
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$,
Model 2T : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(TextMatr_t) + e_{t+1}$.

Table A.9: Squared prediction errors for Ridge Regression, Kernel Ridge Regression, and Lasso Regression with Feature Selection at 1%. Forecast Years: 1999-2010, all firms.

Panel A: Estimation Methodology - Ridge Regression and Feature Selection at 1%.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.007	0.040	0.006
SPE_T-*adj*	0.039	0.006	0.038	0.005
AI	0.002^{***}	0.001^{+++}	0.002^{***}	0.001^{+++}
Statistics	(19.53)	(29.3)	(14.5)	(28.7)
Econ Improve$^{\diamond}$	$54.7^{\dagger}$/36.9		$54.5^{\dagger}$/37.2	

Panel B: Estimation Method - Kernel Ridge Regression and Feature Selection at 1%.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_T-*adj*	0.039	0.005	0.038	0.004
AI	0.002^{***}	0.001^{+++}	0.002^{***}	0.001^{+++}
Statistics	(18.4)	(23.8)	(13.8)	(23.4)
Econ Improve$^{\diamond}$	$52.2^{\dagger}$/37.0		$51.5^{\dagger}$/36.8	

Panel C: Estimation Method - Lasso Regression and Feature Selection at 1%.

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.006	0.040	0.005
SPE_T-*adj*	0.038	0.004	0.038	0.004
AI	0.003^{***}	0.002^{+++}	0.002^{***}	0.001^{+++}
Statistics	(7.51)	(19.1)	(6.65)	(17.8)
Econ Improve$^{\diamond}$	$52.6^{\dagger}$/40.8		$52.0^{\dagger}$/40.9	

SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.
SPE_T-*adj* is the adjusted squared prediction error from the T-model, calculated as the squared difference between actual and predicted *ROE* and adjusted for the squared difference in ROE predictions from the Q- and T-models.
AI is the accuracy improvement gained from using a larger text-enhanced model, calculated as the difference in SPE_Q and SPE_T-*adj*.
$^{\diamond}$ Percentage of observations for which text improves forecast accuracy by 5% or more (first number) and the percentage of observations for which text reduces forecast accuracy by 5% or more (second number).
† Proportion of observations with improved accuracy exceeds the proportion of observations with reduced accuracy using the binomial test at the 1% significance level.
***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test). Number of observations: 23,976.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$,
Model 1T : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(TextMatr_t) + e_{t+1}$,
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$,
Model 2T : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(TextMatr_t) + e_{t+1}$.

Table A.10: Determinants of text informativeness: 1999-2010.

	Predicted Sign	(1)	(2)
ABS_CURR_CH	+	-0.027	-0.028*
		(-1.59)	(-1.65)
ABS_FUT_CH	+	0.938***	0.950***
		(11.28)	(11.40)
CURR_INCR	+	0.031***	0.026**
		(3.10)	(2.56)
FUT_INCR	?	-0.089***	-0.091***
		(-7.43)	(-7.63)
ACCR	+	0.154***	0.148***
		(4.09)	(3.97)
SIZE	+	0.012***	0.011***
		(3.49)	(3.24)
MTB	+	0.002	0.001
		(1.02)	(0.78)
Z-SCORE	-	-0.002*	-0.003**
		(-1.83)	(-2.12)
INTAN	+	0.007	0.016
		(0.74)	(0.58)
BIG4	+	0.009	0.010
		(0.60)	(0.71)
NBSEG	+	0.001	0.002
		(0.13)	(0.27)
NGSEG	+	-0.013*	-0.013*
		(-1.78)	(-1.80)
FOG	-	-0.001	-0.003
		(-0.81)	(-1.25)
LENGTH	-	-0.009	-0.013
		(-0.95)	(-1.29)
S&P_QR	+	0.141***	0.146***
		(9.01)	(9.27)
POST	+		0.047***
			(3.70)
CRISIS	-		-0.046***
			(-2.93)
CDSD	?		-0.031**
			(2.59)
CASH	?		0.057
			(1.42)
CRISIS$\times$ *CDSD*	-		-0.049**
			(-2.31)
CRISIS$\times$ *CASH*	+		0.146**
			(2.02)
Observations		16,658	16,658
$Adj.R^2$		7.12%	7.46%

Dependent variable: *AI*(Q-T) is the pair-wise difference (accuracy improvement) in APE between quantitative and text-enhanced models multiplied by 100. Independent variables: *ABS_CURR_CH* is the absolute change in current earnings; *ABS_FUT_CH* is the absolute change in future earnings; *CURR_INCR* is 1 if $ROE_t > ROE_{t-1}$ and 0 otherwise; *FUT_INCR* is 1 if $ROE_{t+1} > ROE_t$ and 0 otherwise; *ACCR* is earnings minus cash flow from operations scaled by the average book value of owners' equity; *SIZE* is the logarithm of firm's market value of equity; *MTB* is the market value of equity scaled by the book value of total assets; *Z-SCORE* is the Altman z-score; *INTAN* is the intangibles divided by total assets; *BIG4* is 1 if company is audited by BIG4 auditing company (or BIG5 in earlier years of our sample); *NBSEG* is the logarithm of 1 plus the number of business segments; *NGSEG* is the logarithm of 1 plus the number of geographic segments; *FOG* is the Fog index of the MD&A; *LENGTH* is the logarithm of total words in MD&A; *S&P_QR* is 1 if S&P's Quality Ranking of a company is B+ or higher and 0 otherwise; *POST* is 1 if the forecasting period is 2003-2010 and 0 otherwise; *CRISIS* is 1 if the forecasting period is 2007-2009 and 0 otherwise; *CDSD* is 1 if a company belongs to the consumer discretionary sector and 0 otherwise; *CASH* is measured as cash and cash equivalents divided by total assets.
***, **, * indicate significance at the 1%, 5%, and 10% levels, respectively. Standard errors are adjusted for clustering by firm.

Table A.11: Mean, median, and pair-wise differences in *APE* of Q- and T- models, by sub-periods.

	Model 1		Model 2	
	Mean	Median	Mean	Median
PRE 1999-2002				
SPE_Q	0.043	0.007	0.044	0.007
SPE_T*-adj*	0.042	0.007	0.043	0.006
AI(Q-T)	0.001^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(8.42)	(8.46)	(4.07)	(8.02)
Econ Improve$^{\diamond}$	$12.6^{\dagger}$/10.1		$12.6^{\dagger}$/10.2	
POST 2003-2010				
SPE_Q	0.040	0.005	0.039	0.005
SPE_T*-adj*	0.038	0.004	0.038	0.004
AI(Q-T)	0.002^{***}	0.001^{+++}	0.002^{***}	0.001^{+++}
Statistics	(15.53)	(23.06)	(14.81)	(21.66)
Econ Improve$^{\diamond}$	$38.5^{\dagger}$/26.5		$38.1^{\dagger}$/26.5	
POST - no crisis				
SPE_Q	0.033	0.005	0.032	0.004
SPE_T*-adj*	0.031	0.004	0.031	0.003
AI(Q-T)	0.002^{***}	0.001^{+++}	0.002^{***}	0.002^{+++}
Statistics	(13.55)	(23.28)	(13.21)	(22.38)
Econ Improve$^{\diamond}$	$26.1^{\dagger}$/16.0		$25.8^{\dagger}$/15.9	
POST - crisis				
SPE_Q	0.051	0.006	0.051	0.006
SPE_T*-adj*	0.050	0.006	0.050	0.005
AI(Q-T)	0.001^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(7.81)	(7.45)	(7.07)	(6.30)
Econ Improve$^{\diamond}$	$12.4^{\dagger}$/10.5		$12.2^{\dagger}$/10.6	

SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.
SPE_T*-adj* is the adjusted squared prediction error from the T-model, calculated as the squared difference between actual and predicted *ROE* and adjusted for the squared difference in ROE predictions from the Q- and T-models.
$^{\diamond}$ Percentage of observations for which text improves forecast accuracy by 5% or more (first number) and the percentage of observations for which text reduces forecast accuracy by 5% or more (second number).
† Proportion of observations with improved accuracy exceeds the proportion of observations with reduced accuracy using the binomial test at the 1% significance level.
***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test). Number of observations: 23,976.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$
Model 1T : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(TextMatr_t) + e_{t+1}$
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$
Model 2T : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(TextMatr_t) + e_{t+1}$

Table A.12: Accuracy of text-enhanced models and 9-month analyst consensus forecasts.

Panel A: Mean, median, and pair-wise differences in Squared Errors, no analyst following.[†]

	Model 1		Model 2	
	Mean	Median	Mean	Median
SPE_Q	0.041	0.007	0.041	0.007
SPE_T-*adj*	0.040	0.006	0.040	0.006
AI(Q-T)	0.001^{***}	0.001^{+++}	0.001^{***}	0.001^{+++}
Statistics	(7.83)	(7.19)	(6.85)	(6.83)

Panel B: Mean (Median) Squared Error of text (T) and analysts (A) by analyst following.[††]

	Model 1			Model 2		
	High	Medium	Low	High	Medium	Low
SPE_T-*adj*	0.041	0.038	0.039	0.041	0.038	0.038
	(0.007)	(0.006)	(0.006)	(0.007)	(0.006)	(0.005)
SPE_A	0.040	0.039	0.041	0.040	0.039	0.041
	(0.006)	(0.006)	(0.007)	(0.006)	(0.006)	(0.007)
AI(T-A)	0.001^{**}	-0.001	-0.002^{***}	0.001	-0.001^{**}	-0.003^{***}
	(0.001^{++})	(0.001)	(-0.001^{+})	(0.001)	(-0.000)	(-0.002^{++})
t-statistics	2.10	-1.18	-3.09	0.48	-2.22	-3.82
z-statistics	(2.35)	(0.82)	(-1.82)	(1.50)	(-0.95)	(-2.40)

Panel C: Mean (Median) Squared Error of text (T) and analysts (A) by size.[††]

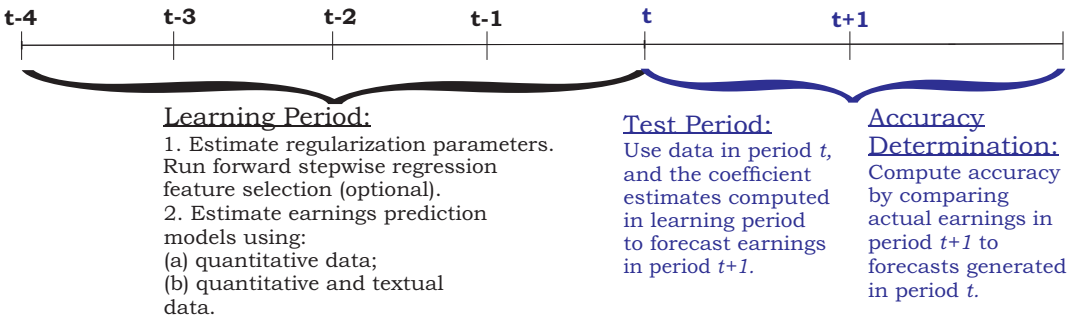
	Model 1			Model 2		
	Large	Medium	Small	Large	Medium	Small
SPE_T-*adj*	0.039	0.040	0.039	0.039	0.039	0.038
	(0.006)	(0.006)	(0.006)	(0.006)	(0.005)	(0.005)
SPE_A	0.036	0.038	0.040	0.036	0.038	0.040
	(0.004)	(0.006)	(0.007)	(0.004)	(0.006)	(0.007)
AI(T-A)	0.003^{***}	0.002	-0.001^{**}	0.003^{***}	0.001	-0.002^{***}
	(0.002^{+++})	(0.001)	(-0.001^{++})	(0.002^{+++})	(-0.001)	(-0.002^{++})
t-statistics	3.55	1.38	-2.14	2.89	1.22	-3.38
z-statistics	(3.87)	(0.82)	(-2.46)	(3.50)	(-0.65)	(-2.35)

SPE_Q is the squared prediction error from the Q-model, calculated as the squared difference between actual and predicted *ROE*.
SPE_T-*adj* is the adjusted squared prediction error from the T-model, calculated as the squared difference between actual and predicted *ROE* and adjusted for the squared difference in ROE predictions from the Q- and T-models.
AI(Q-T) is the accuracy improvement (difference between squared errors) of Q-model and T-model; *AI*(T-A) is the accuracy improvement (difference between squared errors) of T model and analysts consensus forecast (analyst consensus forecast is calculated as analyst forecasts of earnings per share reported on I/B/E/S nine months before fiscal year-end, divided by the beginning-of-year average book value of equity per share).
***, **, * ($^{+++,++,+}$) indicate significance at the 1%, 5%, and 10% levels, respectively, using the one-sided t-test (Wilcoxon signed rank test).
[†] Number of observations: 10,924.[††] Number of observations: 13,052.

Model 1Q : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + e_{t+1}$
Model 1T : $ROE_{t+1} = \beta_0 + \beta_1 ROE_t + \alpha(TextMatr_t) + e_{t+1}$
Model 2Q : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + e_{t+1}$
Model 2T : $ROE_{t+1} = \beta_0 + \beta_1 OPINC_t + \beta_2 NOPINC_t + \alpha(TextMatr_t) + e_{t+1}$

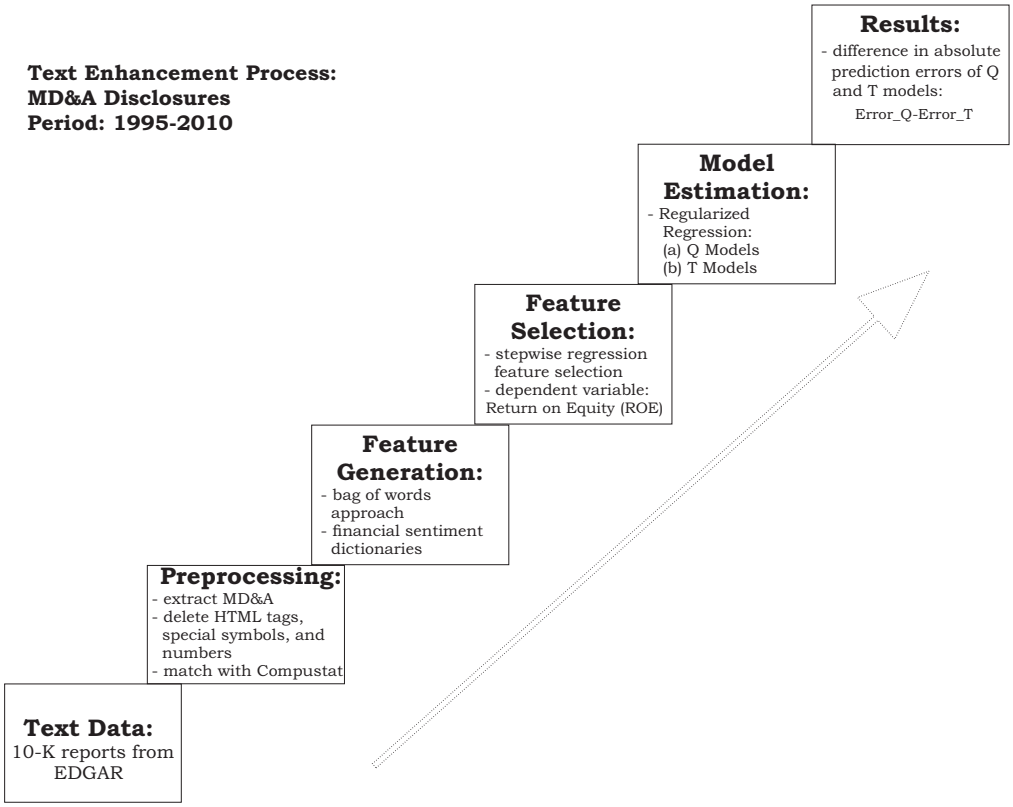
Figures

Figure A.1: Forecasting Timeline



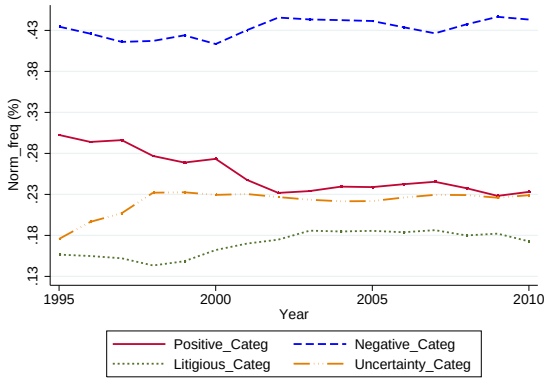
This figure shows the forecasting timeline. First four years, $(t-4), ..., (t-1)$ are used to select relevant variables and estimate the coefficients. Year t is used to make the out-of-sample forecasts of future earnings. Finally, year $t+1$ is used to determine the forecast accuracy of models.

Figure A.2: Text extraction and estimation procedures.

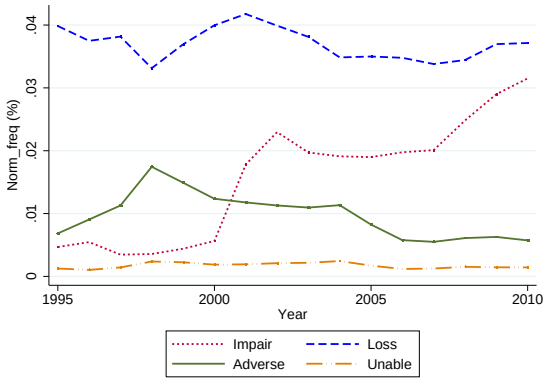


This figure shows the sequence of steps undertaken in the study: a short description of data sources, preprocessing steps, feature generation and selection steps, model estimation and analysis steps. Sample period: 1995-2011. Primary data sources: EDGAR and COMPUSTAT.

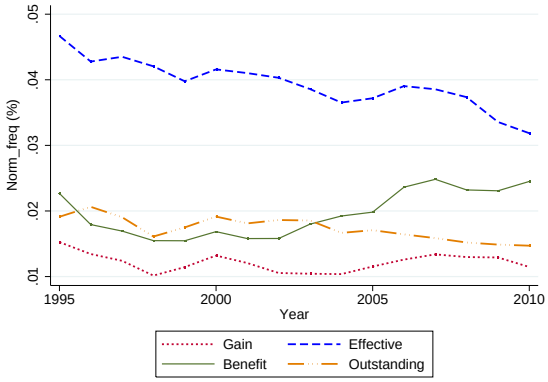
Figure A.3: Mean normalized frequencies: categories and individual words.



(a) Mean normalized frequencies of *Positive*, *Negative*, *Litigious*, and *Uncertainty* Categories

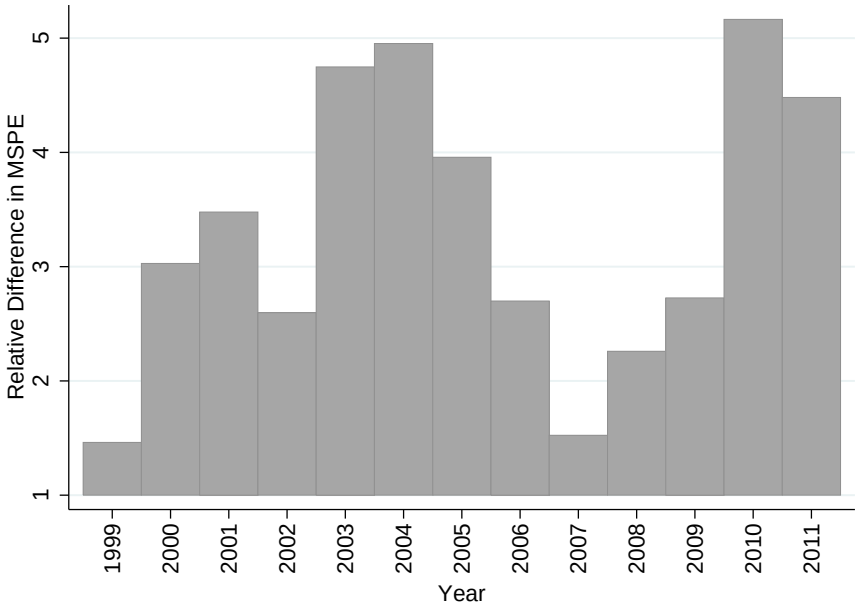


(b) Mean normalized frequencies of negative words *Impair, Loss, Adverse*, and *Unable*



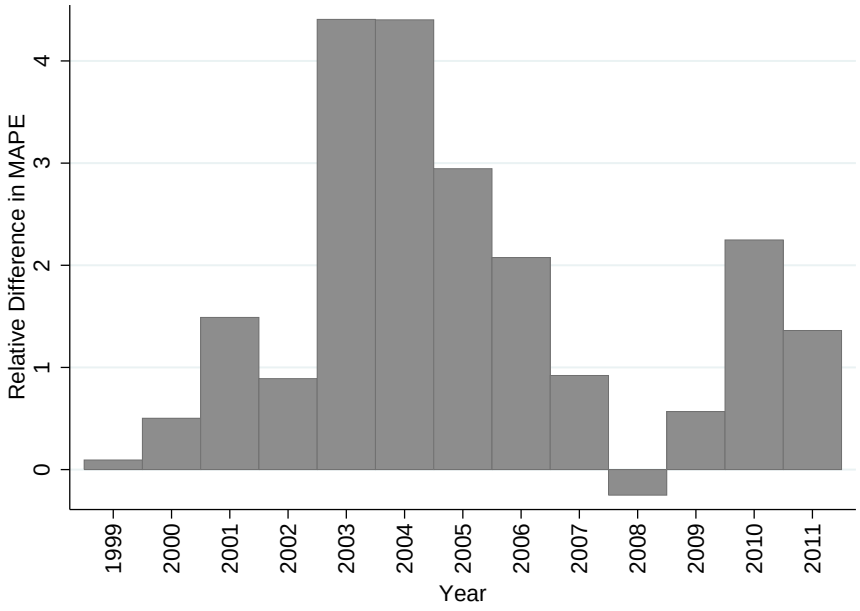
(c) Mean normalized frequencies of positive words *Gain, Effective, Benefit*, and *Outstanding*

Figure A.4: Relative Difference in Mean Squared Prediction Errors of Q- and T- models, year-by-year



This figure plots relative percentage differences in mean squared prediction errors between text-enhanced and quantitative models from 1999-2010, where percentage difference is computed as $(MSPE_Q - MSPE_T)/MSPE_Q$ and multiplied by -1. Positive values indicate text-enhanced model superiority; negative values indicate quantitative model superiority.

Figure A.5: Relative Difference in Mean Absolute Prediction Errors of Q- and T- models, year-by-year



This figure plots relative percentage differences in mean absolute prediction errors between text-enhanced and quantitative models from 1999-2010, where percentage difference is computed as $(MAPE_Q - MAPE_T)/MAPE_Q$ and multiplied by -1. Positive values indicate text-enhanced model superiority; negative values indicate quantitative model superiority.

Curriculum Vitae

Khrystyna Bochkay

1987	Born in L'viv, Ukraine.
1994 - 2004	Attended Primary and Secondary School # 64 in L'viv, Ukraine.
2004 - 2009	Attended Ivan Franko National University of L'viv, Ukraine.
2008	Bachelor of Mathematics, Ivan Franko National University of L'viv, Ukraine.
2009	Master of Science in Statistics, Ivan Franko National University of L'viv, Ukraine.
2009 - 2014	Attended Rutgers University, Newark, NJ.
2014	PhD in Accounting, Rutgers University, Newark, NJ.
2014 - present	Employed by the University of Miami, Coral Gables, FL - Assistant Professor of Accounting.